Informationsvidenskab

1978 kompendium for sektion II/ 2 del

Hjørland, Birger

Publication date:
1977

Document version
Også kaldet Forlagets PDF

Citation for published version (APA):
Hjørland, B. (1977). *Informationsvidenskab: 1978 kompendium for sektion II/ 2 del*. [Danmarks Biblioteksskole].

I N F O R M A T I O N S V I D E N S K A B .

1978 kompendium for sektion II/ 2.del
ved Birger Hjørland

København: Danmarks Biblioteksskole, 1977.

1: Side 1-82

INDHOLD:

Forord.

Det sidste element i uddannelsen til bibliotekar ved forskningsbibliotekerne består af et 4o timers kursus i informationsvidenskab, der afholdes af afdelingen for referencearbejde og dokumentation. De sidste tre år har dette kursus været kørt af Axel Andersen og undertegnede i fællesskab, således at jeg især har taget mig af emnekredsene brugerundersøgelser og informationsgenfinding (samt -søgning), mens Axel Andersen har varetaget bibliometri (litteraturens vækst, forældelse, spredning), den videnskabelige kommunikations udvikling og fremtid m.v.

Oprindeligt anvendtes på dette kursus Mikhailov & Giljarevskij's lærebog (som Chr. Herman Jensen har æren for at have "opdaget"). Denne lærebog overlapper dog så meget med de øvrige dele af bibliotekaruddannelsen, at den ikke i sig selv udgør et egnet lærebogsmateriale, selvom den indeholder virkeligt gode elementer, der derfor også indgår i dette kompendium. Det er vanskeligt at finde en egnet lærebog af to grunde: 1) udbuddet er meget sparsomt, og det meste af udbuddet svarer ikke til hvad vi normalt forstår ved informationsvidenskab, men er snarere f.eks. introduktioner til edb-baseret litteratursøgning. 2) Forskellige elementer af informationsvidenskaben indgår allerede i forskellige fag (f.eks. klassifikation og edb) og undervisningen i informationsvidenskab må derfor tilpasses studiets helhed, hvilket man ikke kan forvente at en lærebog fremstillet til et andet formål vil gøre. I denne situation må idealet være selv at skrive en lærebog, og indtil dette er gjort da køre på enkeltkapitler og -artikler m.v.

Det er af flere forskellige grunde en nødvendighed at have en samling på de benyttede artikler m.v. F.eks. har der fra de studerendes side været et stærkt ønske/krav om at kunne få udleveret materialet i god tid inden selve kursets begyndelse, således at de kunne benytte praktiktid m.v. til at studere materialet, ligesom der har været ønske om at få en introduktion til emnet allerede ved starten af 2.del, så man i forbindelse med andre fag (især specialreference) kan se hvilke emner, der senere vil blive taget op. Der har også været stor efterspørgsel efter det benyttede materiale i forbindelse med arbejde i fagudvalg, studienævnets planlægningsgruppe for informationsvidenskab på sektion I m.v. Endelig er en vis samling meget nødvendig af hensyn til koordineringen af undervisningen fagene imellem, så andre fags lærere kan se hvilke emner, der bliver taget op i dette afsluttende kursus. Dette er begrundelsen for at jeg nu har samlet dette kompendium (eller rettere denne antologi). Det er dog ikke tanken, at undervisningen skal være bundet loo% af kompendiet. I samarbejde med de studerende kan elementer tilføjes, fjernes eller erstattes med andet.

Om rationalen for udvælgelsen af materialet følgende:

Artiklen "Psykologi og informationsvidenskab" har jeg valgt som introduktion fordi den kort slår nogle helt elementære ting fast m.h.t. hvad informationsvidenskab er og ikke er. Den forsøger også at slå bro over de mere tekniske aspekter omkring informationssøgning m.v. og de mere samfundsmæssige aspekter omkring brugerundersøgelser o.lign., og kan derfor forhåbentligt give en sammenhæng mellem undervisningens hovedtemaer, så disse ikke blot står som uafhængige, mere eller mindre tilfældige dele. Iøvrigt kan artiklen betragtes som en filosofisk eller videnskabsteoretisk introduktion til informationsvidenskaben.

Artiklen "Edb-baseret referenceservice" er skrevet for førsteårsstuderende, men medtaget her fordi der vil gå et par år før de studerende på 2.del vil kende artiklen fra 1.dels undervisningen. Dele af artiklen kan anses for ren repetition af stof fra edb-undervisningen, men selv i de mest tekniske afsnit har jeg forsøgt at fremsætte nogle mere principielle konklusioner (f.eks. vedr. muligheden af at automatisere indexerings- og søgefunktioner). Edb-teknik har i og for sig ikke direkte noget med informationsgenfinding at gøre, idet informationsgenfindingssystemer kan være både manuelle og automatiserede. Men edb-teknikken tydeliggør visse principielle forhold, og kan altså betragtes som en pædagogisk (men ikke en logisk) nødvendig forudsætning for en mere principiel beskæftigelse med genfindingsprincipper.

Mikhailov og Giljarevskijs kapitler om informationsgenfinding udmærker sig ved at være en god pædagogisk præsentation af søgesystemer udfra en principiel informationsvidenskabelig synsvinkel. En vis del af stoffet er repetition af klassifikationsundervisningen - dette gælder især afsnittet "Conventionel information retrieval systems" - men dette stof sættes ind i en fælles begrebsmæssig ramme (eksempler på forskellige informationsgenfindingssystemer). Dette stof er ikke tænkt som "eksamensstof", men kun som en forudsætning for at kunne forstå baggrunden for "moderne" informationsgenfindingsprincipper: deskriptorbaseret søgeteknik, herunder brugen af thesauri. Visse dele af Mikhailov & Giljarevskij's kapitler er ren overlapning af andre dele af kompendiet og i og for sig overflødige. I det hele taget erkendes det blankt, at der er behov for en ajourføring og dansk bearbejdning af disse kapitler, men denne opgave har jeg altså kun haft tid til i det omfang stoffet er indarbejdet i andre kapitler.

"Evaluering af informationsgenfindingssystemer" var fra min side tænkt som et undervisningsmateriale, men jeg har haft visse dårlige erfaringer med det. Min idé var at tage udgangspunkt i et informationsvidenskabeligt problem, se hvordan informationsvidenskaben har forsøgt at løse dette problem, og komme med en konklusion. Samtidig er det et meget vigtigt problem, der er valgt. Christian Herman Jensen har nok ramt noget rigtigt, da han lidt bebrejdende fortalte mig, at dette arbejde kunne jeg anvende som et debatindlæg i et fagligt tidsskrift el. lign., men det var ikke undervisningsmateriale for en grunduddannelse til bibliotekar. Jeg vil dog mene, at dette kun gælder sidste del, hvor konklusionerne kommer frem, mens første del er ret introducerende og i og for sig er en lærebogsindføring i definitioner og begreber. Jeg har valgt at medtage denne sag, men kun tænkt at anvende første del. Men jeg vil gerne give interesserede mulighed for at læse det hele og evt. tage stilling til problematikken.

"ERIC-thesaurus" med tilhørende øvelsesopgaver er simpelt hen opøvning i brug af thesaurus, og knytter sig til sidste del af Mikhailov & Giljarevskij's stof. Eksempler og øvelser er valgt på et biblioteksfagligt materiale, således at det er principper for thesaurr og ikke de særfaglige problemer, der skal gøre sig gældende.

Om brugerundersøgelser har jeg kun medtaget afsnittet fra Forskningsbibliotekernes målsætningsudvalgs betænkning. Tidligere har jeg bl.a. anvendt både kapitlet i Mikhailov & Giljarevskij og min og Annette Skovs artikel "Samfundsinformation" (Biblioteket som informationscentral, 3.udg. 1977). Sidstnævnte er ikke

så relevant for forskningsbiblioteker som for folkebiblioteker, og i det hele taget savnes den helt rigtige artikel om dette emne. Det medtagne stof skal nok suppleres med et eller andet, evt. kan de studerende være med til at afgøre dette.

Sidst er medtaget en kort oversigt over nogle informations-videnskabelige håndbøger og bibliografier. Der er jo ikke specialreference indenfor dette område, og det vil være mærkeligt om bibliotekarer netop i deres eget fag skulle kende mindre til disse hjælpemidler end i andre fag. Imidlertid forekommer det ikke videre interessant for de studerende, og man kan jo selv læse det igennem, og bruge det som opslagsværk hvis/når man får brug for det. Dette er altså heller ikke "obligatorisk".

31. oktober 1977

Birger Hjørland

PSYKOLOGI OG INFORMATIONSVIDENSKAB

Birger Hjørland
Danmarks Biblioteksskole, København

Artiklen behandler nogle forbindelseslinjer mellem faget psykologi på den ene side og faget informationsvidenskab på den anden. Med faget psykologi henvises både til den videnskabelige – og den praksisrettede side.

Hvad er informationsvidenskab?

Der hersker langt fra enighed om, hvad informationsvidenskab er, og hvordan feltet skal afgrænses. Dets genstand er overføring og formidling af information, men selve informationsbegrebet er problematisk. Personligt tilslutter jeg mig de forskere, der definerer information som en effekt af en kommunikationsproces på modtageren, og tager altså afstand fra den dagligdags opfattelse af information som synonymt med budskaber, data, oplysninger eller lignende. Der er forskere, der anlægger en universel definition, dvs. opererer med et informationsbegreb, der er fælles for bl.a. fysikken, biologien, psykologien og sociologien. Jeg mener, at konsekvensen af dette er, at informationsvidenskaben kommer til at beskæftige sig med urimeligt abstrakte forhold, og dermed mister forbindelsen med kerneproblemerne, der vedrører den samfundsmæssigt producerede viden. Den samfundsmæssigt producerede viden er bl.a. den videnskabelige viden, og efter den logiske positivismes doktriner udgør denne form for viden den eneste sande erkendelse. I konsekvens af dette har informationsvidenskaben hidtil været domineret af bestræbelser på at fremme overføringen af den videnskabelige viden. Videnskaberne udgør imidlertid kun en blandt flere former for erkendelse, og også den viden, der produceres af andre end forskere, er særdeles væsentlig. Der er da også voksende tendenser til at beskæftige sig med andre former for viden i informationsvidenskaben (se f.eks. Hjørland & Skov, 1977).

Der er andre videnskabelige discipliner end informationsvidenskaben, der har overføring af information som objekt. Dette gælder således massekommunikationsforskningen, pædagogikken og dele af psykologien.

Disse områder er vigtige grænseområder, og der er store overlapninger, f.eks. vedrørende undersøgelser af hvordan viden spredes, søges og udnyttes. Der er imidlertid også væsentlige forskelle. En af dem er informationsvidenskabens interesse for informationsgenfinding eller "information retrieval", dvs. lagring af data på en sådan måde, at en bruger kan trække netop de data ud, der er relevante for hans problem. På denne måde er brugernes behov, initiativ og aktivitet i højere grad udgangspunktet for informationsvidenskaben end for massekommunikationsforskningen eller pædagogikken. Et andet forhold, der karakteriserer informationsvidenskaben er dens interesse for brugerens adgangsveje til dokumenter. Historisk er informationsvidenskaben således en udvikling fra dokumentationsområdet.

Informationsvidenskab og biblioteker

Informationsvidenskaben har en snæver forbindelse med biblioteker. Bibliotekarer kan imidlertid synes utilbøjelige til at se deres profession som et egentligt fag endsige en videnskab. De ser den snarere som en praksis, der naturligvis kræver en meget omfattende viden (f.eks. vedrørende opslagsbøger og kataloger, alfabetiseringsregler, fremmedsprog, edb-teknik og all-round orientering i forskellige videnskaber etc.). Når man taler om biblioteksforskning, er der altfor sjældent tale om en egentlig biblioteksfaglig forskning (dvs. studier af bibliotekernes funktion o.lign.), men ofte om forskning, der blot udnytter bibliotekernes materialer og øvrige ressourcer. I det (beskedne) omfang en egentlig biblioteksfaglig forskning eksisterer, kan vi opfatte den som en del af informationsvidenskaben. Når den kun udgør en del, er det fordi der eksisterer mange andre informationsformidlende systemer end bibliotekerne, og det ofte er ret tilfældigt, om en bestemt service eller funktion varetages af et bibliotek eller en anden form for institution, f.eks. et dokumentationscenter. For at nævne et enkelt eksempel: Der er idag mange centre ud over verden, der har store bibliografiske systemer som f.eks. Psychological Abstracts liggende på magnetbånd, og som foretager søgninger for forskere og andre brugere. En sådan funktion er helt central i informatikken (selvom man kan diskutere værdien af sådanne tjenester), og udforskningen af sådanne systemer kan naturligvis ikke være afhængig af, om de mere eller mindre tilfældigt er knyttet til biblioteker eller til f.eks. regnecentre.

Der er således to hovedforskelle mellem informationsvidenskab og den traditionelle bibliotekskundskab: 1) *Informationsvidenskaben er*

en videnskabelig disciplin. Ikke så meget i kraft af sine resultater, men fordi dette synspunkt afspejler en empirisk og teoretisk interesse, der adskiller sig fundamentalt fra en (ganske vist krævende og kompliceret) optræning i eksisterende bibliotekstekniske funktioner. 2) Informationsvidenskaben indskrænker sig ikke til at studere biblioteker, men studerer *hele det informationssystem, der står til brugernes rådighed,* og som udover biblioteker f.eks. omfatter andre organisationer, personer, møder, tidsskrifter m.v. I dette system indgår såvel sociale komponenter (mennesker) som tekniske komponenter (kartoteker, edb-udstyr m.v.). Sådanne systemer kaldes derfor sociotekniske systemer.

Informationsteknologien

Det er ofte svært at adskille informationsvidenskaben fra informationsteknologien (især edb-teknikken), men efter min – og vel de fleste psykologers – mening, bør videnskaben være overordnet teknikken, idet den skal klargøre hvilken teknik, der er behov for og hvilke funktioner teknikken reelt tjener. Informationsteknologien har skabt mange betydningsfulde ændringer i de allerseneste år, men udfra min synsvinkel har denne teknologi i nogen grad tilsløret og ikke afdækket de egentligt informationsvidenskabelige problemer. Et eksempel vil belyse denne påstand:

Det er idag muligt for en bruger at gå til en edb-terminal, over telefonen kalde en udenlandsk edb-central, og straks være i direkte forbindelse med mange vigtige fagbibliografiske systemer, herunder Psychological Abstracts fra 1967 til dato. Der kan umiddelbart søges på f.eks. ordet "Human Female" og på ordet "Alcohol drinking patterns" og se hvor mange referencer, der er til hvert af disse. I begge tilfælde vil der være langt flere end der kan anvendes (med mindre hensigten er at lave en bibliografi eller lignende), såsøgningen kan nu kombineres, så brugeren kun får de referencer, der både handler om "Human females" *og* om "Alcohol drinking patterns", altså dokumenter, der må antages at behandle kvinders drikkevaner. Antallet af dokumenter vil nu være mere rimeligt til f.eks. en artikel, men hvis der stadigvæk er for mange, kan der afgrænses f.eks. på kvindens alder (eller på dokumentets alder), på dokumenttype (bog, tidsskriftartikel m.v.), på dokumentets sprog m.m.m. Litteraturhenvisningerne kan umiddelbart skrives ud på skriveterminal, og dermed kan udtrækkes alle oplysninger om dokumentet (incl. dets referat). Man kommer ikke uden om, at denne informationsteknologi er raffineret, og at de frustrationer, der knytter sig til

en søgning i de trykte registre på bibliotekerne i høj grad er skudt i baggrunden. Et tilsvarende manuelt arbejde er særdeles tidsrøvende.

Mit synspunkt er imidlertid, at nok har teknikken betydet væsentlige *praktiske* lettelser, men den store kvalitative bedring har den ikke medført. De registrerings- og søgeproblemer, der er forbundet med psykologiens manglende eksakthed er ikke løst bedre fordi systemet er blevet automatiseret. Efter min mening har man brugt meget betydelige materielle og menneskelige ressourcer på at gøre Psychological Abstracts tilgængelig for edb-baseret søgning, men man har brugt altfor få ressourcer på en mere filosofisk analyse af indexeringsproblemer i psykologi. De søgeord, man især opererer med i Psychological Abstracts er betegnelser for afhængige og uafhængige variable og for den eller de undersøgte organismer. Man lægger sig med andre ord stærkt op ad den positivistiske videnskabstradition. Man kan naturligvis sige, at i og med, at de dokumenter man indexerer følger denne tradition, er det naturligt, at selve bibliografien også gør det, men det betyder store vanskeligheder, hvis man skal betjene brugere, der ikke kan eller vil indordne sig denne tradition. Vi har visse danske erfaringer med dette (Hjørland & Bredo, 1977), og tilsyneladende er problemerne på dette felt større i Danmark end f.eks. i Sverige (se Ness, 1977). Det er min opfattelse, at systemet kunne være lavet bedre, også selvom man accepterer at bibliografien må afspejle litteraturen som den er, dvs. med en stærk positivistisk dominans. Ved udviklingen af Psychological Abstracts' edb-system har man som vanligt i den positivistiske tradition haft naturvidenskaberne som forbillede. Men den registertype, man opererer med, er mere velegnet på f.eks. tekniske og naturvidenskabelige områder. Den stærke vægt på højt specialiserede og frit kombinerbare begrebselementer, sker i et område som psykologien på bekostning af struktur, bredde og sammenhæng. Disse egenskaber repræsenterer andre systemtyper i højere grad, f.eks. de såkaldte facetterede registre, som det her vil føre for vidt at komme ind på. Problemer af denne art er sande informationsvidenskabelige problemer og ikke blot edb-tekniske problemer. At der ikke bare er tale om, at amerikanerne til eget brug udfra informationsvidenskabelige undersøgelser af egne behov har lavet det system, de fandt bedst, fremgår af den kendsgerning, at alle publicerede undersøgelser synes at vise, at Psychological Abstracts heller ikke for amerikanere synes at fungere tilfredsstillende (se f.eks. Katzer, 1973, p. 329).

"Information retrieval"

Jeg har allerede tidligere defineret begrebet "information retrieval".

Betegnelsen kan synes uheldig, idet information hverken kan opbevares eller genfindes, hvis man anlægger den synsvinkel, at information er en effekt på en modtager. Det, der bliver fundet er altså data eller budskaber. Betegnelsen er imidlertid nu godt fastslået i forskningslitteraturen og kan derfor vanskeligt ændres. Selve informationsgenfindingsproblemet er et meget centralt problem både i informationsvidenskaben, i edb-teknikken og i kognitionspsykologien. I sidstnævnte interesserer man sig for, hvordan den menneskelige hjerne organiserer data og finder dem frem igen.

De fleste genfindingssystemer i informationsvidenskaben bygger på sproglige betegnelser, begreber, og på relationer mellem begreber, f.eks. slægt-arts relationer eller del-helhedsrelationer. Psychological Abstracts system er et eksempel på dette. Hovedproblematikken ved sådanne systemer er naturligvis begrebernes betydning eller semantik. Hvis den, der søger information, benytter begreberne i samme betydning som den, der skriver dokumenterne og den, der indexerer dem i bibliografien, så var mange problemer løst. Man forsøger at råde bod på dette ved at stille en thesaurus til brugernes disposition. Det er en normativ ordbog, der øver synonym- og homonymkontrol, og som henviser brugeren til betydningsrelaterede termer (som regel overbegreber, underbegreber og sideordnede begreber). Men sådanne problemer kan i princippet aldrig løses endeligt. En persons sprogbrug er stærkt afhængig af hans viden og holdning, og den veksler derfor fra dag til dag, og fra person til person. Fra psykologisk side er dette emne bl.a. behandlet i Lindsay & Norman, 1972 (specielt s. 431-433), jfr. Spang-Hanssen (1974, s. 20). Problemet udgør alle emneordsbaserede systemers akilleshæl, og jo større uenighed der er i betydningen af begreberne, desto dårligere fungerer systemer, der bygger på sådanne principper.

Der findes en radikalt forskellig mulighed. Den udnyttes i de såkaldte citationsindexer, hvor den, der hedder "Social Sciences Citation Index" og begyndte at udkomme i 1973, vil være mest relevant for psykologer. (Den findes efterhånden på mange af de større forskningsbiblioteker i Norden, men er for kostbart for mindre institutbiblioteker). Den er en tidsskriftindex, der indexerer artikler i flere tusinde samfundsvidenskabelige tidsskrifter, herunder flere hundrede psykologiske. For alle de artikler, der indexeres, gengives artiklens litteraturliste. Ved hjælp af edb har man så konstrueret citationsindex, hvor det er muligt at slå op på en bog eller artikel fra litteraturlisten og så se, hvad det er for en artikel, der citerer det pågældende dokument. Man kan altså se hvem der f.eks. i de løbende tidsskrifter citerer Georg

Rasch eller en selv, eller hvad man nu måtte være interesseret i. Eller man kan se hvem der f.eks. både citerer Karl Marx og Sigmund Freud, og så gå ud fra, at de pågældende arbejder tilhører den freudo-marxistiske retning. Filosofien er, at der mellem en artikel og den litteratur, som denne artikel citerer, er en emnemæssig forbindelse, og at denne forbindelse kan udnyttes i informationsgenfinding. Ligesom de systemer, der bygger på emneord, har imidlertid også citationsindexerne deres akilleshæl: Det er en forudsætning, at der er en god citeringspraksis, at forfattere kun citerer værker, der har betydning for den pågældende problemstilling, og også – men mindre kritisk – at forfatterne har foretaget en nogenlunde god litteratursøgning, så de kender de værker, der har interesse. I praksis supplerer emneordssystemer og citationsindexer hinanden. Det første vil være bedre til f.eks. at få overblik over alle opfattelser af et begreb (f.eks. psykoterapi), det sidste vil være mest velegnet til at optrævle en bestemt skole, der ikke er repræsenteret ved et særligt begreb.

Ovenstående er et forsøg på at vise, hvordan man til i dag har forsøgt at løse genfindingsproblemerne i informationsvidenskaben. Hvis vi sammenligner disse systemer med den beskrivelse, som kognitionspsykologerne giver af den menneskelige hjernes funktion, vil vi se, hvor langt der er igen, før vi har udviklet "intelligente" systemer. En spændende og lettilgængelig behandling af dette problem giver Lindsay & Norman (1972, kap. 10 og 11) i bogen Human Information Processing. De viser, at problemet med sådanne systemer ikke er at få data ind i systemet (som f.eks. mange bibliotekarer er tilbøjelige til at mene), men at kunne finde dem frem igen. Den menneskelige hukommelse er ikke (som f.eks. biblioteker og edb-maskiner) blot en passiv opbevaring af data, men informationsgenfindingsprocessen i hukommelsen er et spørgsmål om problemløsning. Det er f.eks. meget betydningsfuldt, at et menneske kan vurdere, hvorvidt det er muligt eller sandsynligt at finde svaret på et givent spørgsmål i hukommelsen *inden* en tidsrøvende søgning påbegyndes. Et væsentligt princip, der adskiller den menneskelige hukommelse fra en edb-maskines er, at en menneskelig dataenhed ikke opbevares isoleret, men ændres som følge af nye data. Dette hindrer udsendelsen af store mængder forældet stof, og ville være et overordentligt væsentligt problem at få løst ved de tekniske systemer.

Der er gjort forsøg på at opbygge informationsgenfindingssystemer direkte på kognitionspsykologiske principper, bl.a. på grundlag af Piaget og Gagné (Neill, 1975; Farradane, 1975; m.fl.). Jeg er imidlertid noget skeptisk overfor en umiddelbar og direkte anvendelse. Jeg mener ube-

tinget, at psykologien kan øge vor forståelse af sådanne processer og dermed præge vor holdning, men jeg har ikke set nogen overbevisende demonstration af psykologiens egnethed som direkte rettesnor for udformningen af sådanne systemer. Alle systemer, der bygger på principper fra "menneskets natur" eller "samfundets struktur" kan være velegnede til gårsdagens spørgsmål. Men viden udvikles jo netop ved at der skabes nye relationer mellem emner og derfor vil organisering af data efter bestående psykologiske, lingvistiske eller sociologiske strukturer være uegnet netop til besvarelsen af egentligt nye spørgsmål. Psykologiens rolle bliver derfor hovedsagelig af ergonomisk art, og retningslinier for organisering af data må snarere søges i erkendelses- og videnskabsteori, sådan som det antydedes ved omtalen af Psychological Abstracts.

Interview- og spørgsmålsformulering

Et meget centralt "ergonomisk" problem for informationsvidenskaben er spørgsmålsformuleringen hos brugeren og dialogen mellem brugeren og "systemet", herunder mellem brugeren og en intermediær, f.eks. en bibliotekar. Det er et område, hvor psykologien allerede har ydet et godt bidrag, og hvor der er et oplagt behov for yderligere forskning. Psykologen og informationsspecialisten Robert S. Taylor (1968) har ydet et meget væsentligt bidrag til disse problemer. Han viser, hvorledes man kan opdele brugerens spørgsmålstilblivelse i faser fra det helt ubevidste behov over et bevidst ønske om en bestemt oplysning, til et spørgsmål, der rummer et kompromis mellem den ønskede oplysning og hvad han forventer at bibliotekaren (der jo ikke kan være så godt inde i det pågældende problem, som han selv) kan levere. Brugere tenderer mod at formulere spørgsmål i for brede og overordnede kategorier, og ved at lære bibliotekaren interviewteknik og lånerpsykologi kan man bedre få brugeren til at formulere spørgsmål på en måde, der kan danne udgangspunkt for informationssøgning.

Brugerbehov og -adfærd

Der findes nogle områder indenfor biblioteksvæsnet, hvor behovet for psykologisk ekspertise er åbenlyst, og også højt prioriteret af udøverne. Det drejer sig f.eks. om børnebiblioteksområdet, hvor kendskab til udviklingspsykologi og til børns forhold til forskellige medier er basalt for at kunne udøve funktionen som bibliotekar. Det er i denne forbindelse

bemærkelsesværdigt, at hvor folkeskolen har en udbygget psykologisk tjeneste, megen psykologi i folkeskolelæreruddannelsen og mulighed for videreuddannelse på dette område, da mangler dette aspekt næsten helt i biblioteksuddannelsen. Et andet område er litteraturvidenskab, hvor lånernes oplevelser og udbytte af litteratur (især skønlitteratur) vurderes med henblik på bibliotekarens vejledende funktion. Jeg har ikke personligt beskæftiget mig meget med dette område, men der findes efterhånden en del litteraturpsykologisk faglitteratur, og den russiske bibliotekar og psykolog Nicholas Rubakin (Simsova, 1968) har opstillet en hel bibliotekspsykologi, hvor lånerens forhold til litteraturen og metoderne til at undersøge dette forhold spiller en central rolle. Disse to områder for psykologien henregnes almindeligvis ikke under informationsvidenskaben. Det første måske fordi man ikke almindeligvis betragter børn som beslutningstagere, og hermed som havende informationsbehov (f.eks. i betydningen data til brug i en beslutningsproces), hvad der måske kan være rigtigt i den forstand, at voksne ofte træffer beslutninger på børns vegne, men hvad der i al fald ikke kan anses for acceptabelt. Det andet, nemlig litteraturvidenskab, henregnes ikke under informationsvidenskaben, idet man noget forenklet har skelnet mellem information, viden og fakta på den ene side, og fiktion eller "kultur" på den anden.

Ligesom man kan interessere sig for børn som brugergruppe, kan man interessere sig for f.eks. studerende, ufaglærte arbejdere, forskere etc. Medens det er ret indlysende, at bibliotekaren må have kendskab til udviklingspsykologi, kan det måske være vanskeligt at se, at hun også skal have kendskab til f.eks. forskeres behov og informationssøgningsadfærd. Det er almindeligt at betragte sidstnævnte gruppe udfra et rationelt-økonomisk menneskesyn, hvad der f.eks. kan resultere i, at man overvurderer forskerens ihærdighed og tålmodighed ved f.eks. at prioritere et stort materialeudbud på bekostning af en lettere tilgængelighed til samme. Gentagne undersøgelser i mange fag viser imidlertid, at dette er fejlagtigt: videnskabsmænd og teknikere har som regel tilgængeligheden som det afgørende kriterium, når en informationskilde vælges, og dens kvalitet og alsidighed tages kun i betragtning hvis kilden forekommer tilgængelig rent psykologisk (f.eks. Gerstberger & Allen, 1968). Hermed definerer to klare opgaver sig for psykologer: 1) At kortlægge barrierer mellem brugeren og de informationskilder, der står til hans rådighed, 2) at skabe feed-back mekanismer mellem brugeren og de, der professionelt beskæftiger sig med information. Parker & Paisley (1966) opstiller foruden disse to opgaver en tredie: at opstille

kriterier for udformning af systemer på baggrund af undersøgelser af brugernes behov og evaluering af eksisterende systemers funktion.

Der er knyttet mange vanskeligheder og begrebsmæssige uklarheder til disse forskningsopgaver. Det må for det første understreges, at brugere er et dårligt udtryk, idet ikke-brugere eller potentielle brugere er ligeså vigtige som aktuelle brugere; for det andet, at bruger*behov* ikke er det samme som brugernes *ønsker* eller *krav*. Generelt er brugernes viden om de tilgængelige informationskilder nemlig så dårlig, at de ikke på rimeligt grundlag kan udtale sig om deres egne behov. Disse forhold er den eksisterende forskningslitteratur sig dog bevidst, selvom kun de bedste undersøgelser rent faktisk søger at tage højde for dem. Det, der derimod i højere grad er en svaghed ved brugerundersøgelserne (og hvor disse f.eks. er ubehjælpeligt bag efter massekommunikationsforskningen), er i de videnskabs-teoretiske forudsætninger. Forskningen har således hidtil hvilet på positivistiske og mekanisistiske metoder. En konsekvens af denne videnskabstraditions dominans i brugerundersøgelserne er, at man forsøger at fragmentere adfærden og opstille afhængige og uafhængige variable indenfor områder som modtagerkarakteristika, budskaber, kanaler m.v. Vickery (1973, 43-45) har ligefrem opstillet en hel tabel, der opsummerer de sammenhænge mellem sådanne variable, som forskellige brugerundersøgelser er nået frem til. Men i og med, at man således helt skærer den sociale kontekst væk, som den pågældende adfærd indgik i, mister resultaterne det meste af deres værdi. Der er stærkt brug for beskrivende undersøgelser, der medinddrager sociale determinanter, og som ikke blot "psykologiserer".

Tre modeller for vidensudnyttelse

Havelock (1973) har foretaget en omfattende gennemgang af og konklusion på den litteratur indenfor psykologi, sociologi m.v., der omhandler udnyttelsen af ny viden. Han finder, at der i litteraturen er tre begrebsmæssige modeller. og karakteriseringen af disse kan være oplysende. Den første model kalder han *"forsknings- udviklings- og diffusionsperspektivet"*, dvs. det rationelt tekniske approach. Denne model opfatter overføring af information som en rationel proces, der forløber i klare led, som må omfatte en stærk arbejdsdeling, og forudsætter en passiv bruger. Jeg vil mene, at edb-holdningen til informationsproblemet er et typisk eksempel på denne model. Den anden model hedder *"social-interaktions-perspektivet"*. Herunder henfører Havelock en i hovedsagen sociologisk disciplin, der interesserer sig for, hvordan en

konkret nyskabelse, "innovation" (f.eks. et nyt lægemiddel) rent faktisk spredes, og hvad det er for sociale kræfter, der påvirker denne spredning (uformelle kontakter) sociale strukturer m.v.). Det er en stærkt empirisk funderet retning, som måske er noget svag når det gælder konkrete strategier for informationsformidling. Tilhængere af det sociale interaktionsperspektiv er imidlertid generelt skeptiske overfor hensigtsmæssigheden af forsknings-, udviklingsperspektiver, idet de mener, at det helt overvejende er sociale kræfter, der styrer informationsoverføringen, og at opbygningen af systemer, der ikke tager hensyn til sådanne sociale strukturer er dømt til at fejle. Konsekvensen af dette perspektivs resultater må være at fremme den uformelle kommunikation f.eks. ved møder, henvendelse til "opinion leaders" etc. Den tredie model hedder *"problemløserperspektivet"* og Havelock karakteriserer dette perspektiv udfra fem teser: 1) Brugeren er udgangspunktet. Hvor f.eks. social-interaktionsperspektivet tager en innovation ad notam, understreger problemløserperspektivet, at innovation er meningsløs uden brugerens behov. Havelock bemærker, at denne tese af objektivistisk indstillede forskere betragtes som et moralsk-etisk spørgsmål, der ligger udenfor videnskabens domæne. 2) Diagnose går forud for løsningsforslag. 3) Hjælpen udefra skal være ikke-dirigerende, dvs. brugeren skal have vejledning og træning i at klare *sin* egen problemløsning. 4) Der skal lægges stærkt vægt på brugerens indre ressourcer i forhold til viden udefra. 5) De ændringer, der er iværksat af brugeren selv, vil være de stærkeste og vare længst. Derfor må brugeren internalisere innovationen.

Hvad er konsekvenserne for psykologien af disse tre modeller? Svaret er delvis givet med karakteriseringen af modellerne, men skal for tydelighedens skyld fremhæves.

Forsknings-, udviklings- og diffusionsperspektivet udmærker sig først og fremmest ved at være en ikke-psykologisk måde at betragte informationsformidlingen på. Det er her et spørgsmål om teknisk-økonomisk rationalitet. Det betyder ikke, at psykologien ikke spiller nogen rolle. De psykologiske forhold, der anførtes under informationsgenfinding vil naturligvis være af interesse, bl.a. hele ergonomien (menneske-system tilpasningen). Tidlige arbejdspsykologiske teorier som "scientific management" passer ind i denne model. Der vil også være behov for brugerundersøgelser, men brugeren ses da som et passivt objekt, der er det svage led i et ellers perfekt informationssystem. I det store og hele er psykologien reduceret til en hjælpedisciplin efter dette approach.

I social-interaktionsperspektivet er psykologens rolle mere betydende. Det meste af socialpsykologien kan anvendes på informationsoverføringsproblemer. Schmuck (1968) giver et eksempel på en analyse af de socialpsykologiske kræfter, der stiller sig i vejen for at skoleinspektører kan anvende adfærdsvidenskabelig viden i deres arbejde. Han redegør for de holdninger og fordomme, der gør sig gældende mellem de pædagogiske forskere på den ene side og skolefolkene på den anden. Hvordan manglende face-to-face kontakt fremmer in-gruppe og out-gruppefølelser, ledsaget af gensidige stereotypier, lavt tillidsforhold etc. Altså anvendelsen af traditionel socialpsykologi. Schmuck beskæftiger sig også med andre informationskanaler, såsom tidsskrifter, møder etc. Benyttelsen af disse er også mangelfuld og underlagt tilsvarende socialpsykologiske forhold. Den kur Schmuck foreslår er bl.a. kurser i interpersonel sensivitetstræning, organisationsudviklingsprogrammer etc. Det, der naturligvis præger Schmucks behandling er mangelen på inddragen af samfundets makrostruktur som determinant for disse gruppers arbejdsvilkår og manglende fælles mål. Jeg skal kort vende tilbage til dette i et senere afsnit.

Problemløserperspektivet forekommer mest spændende ud fra en psykologisk synsvinkel. Der er lagt op til en aktionspræget handleform, hvor man tager udgangspunkt i brugernes her-og-nu problemer, hvor man styrker hans selvtillid og hermed evnen til at udnytte såvel sine egne indre ressourcer som nødvendige ydre ressourcer. Det er vanskeligt fra dette perspektiv at adskille informationsproblemer fra andre opgaver, f.eks. psykoterapi og politisk bevidstgørelse. Udfra en informationsbetragtning må det være psykologens opgave at få brugerens til at se, at hans "private" problem også skyldes ydre forhold, som han sammen med andre kan arbejde for at ændre. Måske kan det være svært at se, hvor informationsformidlingen blev af i denne forbindelse. Den amerikanske sociolog og biblioteksskolelærer Mary Lee Bundy (1972) har givet en spændende fremstilling af, hvad bibliotekarens – og hermed informationsformidlerens – rolle er i denne forbindelse. Det er et spørgsmål om at formidle information, der kan løse konkrete sociale opgaver, f.eks. boligproblemer, forbrugerproblemer, sundhedsproblemer etc. I erkendelse af at problemerne hænger sammen, kan psykologer, bibliotekarer, socialrådgivere, jurister, socialmedicinere m.fl. arbejde sammen i rådgivningscentre. En model af denne art er f.eks. anvendt i "Den alternative rådgivning i Immervad (Agger & Richardt, 1977). Ligeledes har den sociale Højskole i København taget et iniativ i samme

retning ved at åbne en "Informationsbutik" i Silkeborggade, som rummer mulighed for et tværfagligt arbejde omkring konkrete livsproblemer, udfra problemløserperspektivet og dermed også informationsperspektiver. Projektet følges af psykologer og bibliotekarer.

Havelock slutter med at præsentere sin egen model, der er et kompromis mellem de tre præsenterede. Dette er måske også, hvad der er behov for. Alle tre modeller indeholder nødvendige og værdifulde elementer. Det er imidlertid spørgsmålet, om der ikke mangler en helt fjerde model? En model, der i højere grad inddrager samfundets makrostrukturer og som forsøger at se ideologikritisk på vidensproduktion og formidling? Havelock nævner (p. 11-19) at Hegel og Marx har "rudimenter til et alternativt perspektiv på informationsspredning og udnyttelse". Men er alle modellerne ikke kun rudimenter?

Samfundsopfattelse og informationsformidlingsteknik

Man kan groft sagt skelne mellem to samfundsopfattelser: *konsensus- og konfliktopfattelsen.* Haslebo (1971) giver en karakteristik af konsensusopfattelsen, som hun betegner den *legitime og monistiske opfattelse* af samfundet. Efter denne opfattelse findes der et centrum for social og politisk magt. Dette centrums magt opfattes både af forskeren og af borgeren som helt legitimt, da centret er i stand til og interesseret i at erkende periferiens behov. Jo mere centret ved om dette behov, desto mere vil det blive til gavn for hele samfundet. Videnskabsmandens opgave vil derfor blive at stille denne viden til rådighed for magtcentret. Haslebo skelner mellem to konfliktopfattelser. Den første går på, at samfundet er *polariseret og udnyttende,* og denne opfattelse indebærer, at samfundet ses som opdelt i en magtelite, der sidder inde med alle magtressourcer, og periferien, der består af masserne, som ikke har adgang til disse ressourcer. Magtelitens interesse er at udnytte periferien enten ved magt eller tvang. Den er interesseret i information, der kan understøtte dens position. Den anden konfliktopfattelse ser samfundet som bestående af mange konkurrerende magtcentre. Med andre ord: samfundet ses som *pluralistisk og konfliktpræget.* Efter disse konfliktopfattelser kan videnskabsmanden ikke arbejde uden at vælge enten at arbejde for magteliten mod periferien eller – efter en diagnose af magtrelationerne – at arbejde for periferien mod magteliten. Han kan f.eks. vælge at informere periferien om at den bliver udnyttet, eller han kan vælge bevidst at modarbejde magtelitens manipulationer.

Haslebos artikel drejer sig om samfundsvidenskabelig forskning i al-

mindelighed. Dens konklusion er, at der er sammenhæng mellem samfundsopfattelse og forskningsmetode. Hvis man har en konsensusopfattelse, kan mekanistiske eller positivistiske metoder være på deres plads, men efter en konfliktopfattelse vil sådanne metoder vække mistillid og forskningsresultaterne blive upålidelige. Organiske eller aktionsforskningsprægede strategier, hvor forskningen foretages i samarbejde med de mennesker, hvis behov undersøges, bliver nødvendige. I og med, at informationsvidenskab er en samfundsvidenskab, kan disse konsekvenser naturligvis overføres til informationsvidenskabens metoder. Konsekvenserne rækker imidlertid langt videre end blot til udforskningen af informationsbehov og -problemer.

Hvis man har en konsensusopfattelse, så arbejder videnskaben i alles interesse og informationsformidlingens opgave bliver så effektivt som muligt at skabe kontakt mellem forskningsresultater (eller andre budskaber) og brugere. Mekanisistiske og edb-mæssige initiativer forekommer indlysende, idet der ikke er nogen grund til at indbygge fagkritiske elementer i kommunikationssystemet. Har man derimod en konfliktopfattelse, så må man opgive tanken om et neutralt informationssystem. Brugerens interesse består da i at afsløre budskabernes (falske) ideologi og identificere dokumenter, der bygger på samme samfundsopfattelse og kæmper for de samme mål som han selv. Dette peger mere i retning af decentrale løsninger og menneskelig fortolkning end verdensomspændende edb-systemer og automatiserede søgefunktioner. Konklusionen må derfor være, at der er en afgjort sammenhæng mellem samfundsopfattelse og informationsstrategi. *Man kan ikke have en informationsvidenskabelig opfattelse uden en samfundsopfattelse.*

Perspektiver

I dette afsluttende afsnit skal jeg pege på nogle perspektiver for informationsvidenskabens udvikling og for psykologiens rolle i denne forbindelse. Det mest afgørende bliver her, om der kan ske en *tilnærmelse mellem teori og praksis.* Der er mange mennesker der udfører informationsarbejde uden nogen teoretisk opfattelse, og der er mange der forsker i informationsproblemer uden at afprøve dem i praksis. Et initiativ som den tidligere omtalte informationsbutik i Silkeborggade er et vigtigt skridt i denne retning.

I teoretisk henseende mener jeg, at de mest værdifulde bidrag i øjeblikket vil være at gå ud fra vesttyskeren Jurgen Habermas' kritiske videnskabsteori og hans offentlighedsteori, der bl.a. er videreudviklet af

Negt & Kluge (1974) og Negt (1975) og forsøge at overføre disse teorier på informationsformidlingens område. Teorierne er velkendte indenfor massekommunikationsforskningen og medieforskningen, men jeg skal her antyde deres anvendelse indenfor informationsgenfindingen (information retrieval).

Der er to hovedproblemer i et informationsgenfindingssystem (som f.eks. det før omtalte Psychological Abstracts): Problemet om bibliografisk kontrol (hvad kommer med i systemet) og problemet om bibliografisk organisering (hvordan opbevares data, så de kan findes igen).

Habermas's metode består i at underkaste videnskaben eller den borgerlige offentlighed en historisk analyse i relation til samfundsstrukturen. Vi kan gøre det samme med den bibliografiske kontrol. Denne opstod oprindeligt i Europa udfra monarkers ønske om at have kontrol med, hvad der blev trykt og formidlet til landets borgere, altså udfra censurinteresser. Senere i samfundsudviklingen, da borgerskabet får den reelle magt, skifter funktionen. I USA bygger den bibliografiske kontrol således på forlæggernes frivillige aflevering af bøger. Deres interesse i bogfortegnelser var hensynet til copyright'en. I videnskabelige informationssystemer som Psychological Abstracts ligger den reelle funktion for den bibliografiske kontrol i at producenterne af psykologisk litteratur ønsker deres arbejder læst – ikke så meget af direkte økonomiske grund, som fordi det for videnskabsmanden er afgørende at få sine dokumenter "anerkendt". *Systemet er således producentorienteret og ikke brugerorienteret,* og selvom de fleste brugere også er producenter, er rollerne og dermed behovene vidt forskellige. Det samme kan siges for den bibliografiske organisations vedkommende. De opslagsord (deskriptorer), der anvendes i registrene er ofte begreber hentet fra dokumenttitler eller -referater. De kan således i princippet opfattes som en slags *reklametermer* for dokumentet, idet de udtrykker hvad producenten gerne vil give brugeren indtryk af, at dokumentet handler om. Alternativet ville være, at sådanne systemer tjente rene brugerinteresser, dvs. kritisk analyserede dokumenter udfra en bestemt brugergruppes synsvinkel og beskrev dokumenterne i brugernes sprog og udfra deres interesser. Marxisten Werner Dube (1975) kalder ironisk amerikanske informationssystemer som Psychological Abstracts for GIGO (Garbage In – Garbage Out), og det vil mange borgerlige videnskabsmænd sikkert give ham ret i.

Hvis vi ser i en bog som Negt (1975): Sociologisk fantasi og eksemplarisk indlæring, så aftegner der sig klare opgaver for, hvordan psykologer kan medvirke til at klarlægge betingelser for politisk bevidsthed. I

informationsvidenskabelig sammenhæng kan disse opgaver overføres til at undersøge informationssystemernes og informationsformidlingens rolle i denne forbindelse. Der må udvikles en "informationsbevidsthed" og skabes forudsætninger for, at "periferien" kan opbygge sine egne informationssystemer.

SUMMARY

Psychology and information science

The article discusses the use of psychology in Information Science problems. Information Science problems should not be confused with purely technical problems. For example the transfer of Psychological Abstracts onto computer is seen as a technical problem, whilst indexing methods are seen as an Information Science problem. It is shown that the field of Human Information Processing can shed light on information retrieval problems, but emphasis is placed primarily on analyses of the theory of knowledge. Amongst psychological applications can be named, amongst others: interview methods; question negotiation; user needs and behavior. Traditional user-studies are criticized for their positivist tendencies and fragmentation of user behavior. Havelock's three models for utilization of knowledge are discussed. Models are requested, which to a greater degree incorporate society's macrostructure. The consequences of a consensus and a conflict model of society for information dissemination are discussed. It is shown that the science of information could with advantage incorporate Jurgen Habermas's critical theories in its foundations.

LITTERATUR

Agger, N.P. & Richard, J: Terapi som politisk praksis og som politisk projekt. *Dansk Psykolognyt*, 1977, *31*, 67-84.

Bundy, M.L.: Urban information and public libraries: A design for service. *Library Journal*, 1972, *96*, 161-169.

Dube, W.: Zur Bildung des Bibliothekars im gegenwärtigen Stadium des amerikanischen Imperialismus. *Zentralblatt fur Bibliothekswesen*, 1975, 89, 199-212.

Farradane, J.E.L.: The psychology of classification. *Journal of Documentation*, 1955, *11*, 187-201.

Gerstberger, P.G. & Allen, T.J.: Criteria used by research and development engineers in the selection of an information source. *Journal of Applied Psychology*, 1968, *52*, 272-279.

Haslebo, G.: Samfundsvidenskabelig målforskning som interventionsform i samfundet. *Nordisk Psykologi*, 1971, *23*, 402-414.

Havelock, R.G.: *Planning for innovation through dissemination and utilization of knowledge*. Ann Arbor: University of Michigan, 1973.

Hjørland, B. & Bredo, O.: *Forsøgsvirksomhed med bibliografiske databaser i psykologi og pædagogik*. København: Danmarks pædagogiske bibliotek. (Mimeo.). Forventes udsendt i 1977. Ca. 100 s.

Hjørland, B. & Skov, A.: Samfundsinformation og biblioteker. I: A. Andersen (red.): *Biblioteket som informationscentral.* 3. udg. København: Danmarks Biblioteksskole, 1977. S. 134-153.

Katzer, J.: The cost-performance of an on-line, free text bibliographic retrieval system. *Information Storage and Retrieval,* 1973, *9,* 321-329.

Lindsay, P.H. & Norman, D.A.: *Human information processing. An introduction to psychology.* New York & London: Academic Press, 1972.

Negt, O.: *Sociologisk fantasi og eksemplarisk indlæring. Til teori og praksis i arbejderuddannelsen.* Roskilde: RUC's boghandel og forlag, 1975.

Negt, O. & Kluge, A.: *Offentlighet og erfaring.* Til organisasjonsanalysen av borgerlig og proletarisk offentlighet. Nordisk Sommeruniversitet, 1974.

Neill, S.D.: Farradane's relations as perceptual discriminations. *Journal of Documentation,* 1975, *31,* 144-157.

Ness, E.: *Forsøk med edb basert litteratursøkning innen psykologi, pedagogikk og specialpedagogikk i Danmark, Finland, Norge og Sverige.* Sammenfattende sluttrapport. Nordisk Utredningsserie, NU 1976:29. Stockholm, NORDDOK, 1977.

Parker, E.B. & Paisley, W.J.: Research for psychologists at the interface of the scientist and his information system. *American Psychologist,* 1966, *21,* 1061-1071.

Schmuck, R.: Social psychological factors in knowledge utilization. I: T.L. Eidell & J.M. Kitchel (Eds.): *Knowledge production and utilization in educational administration.* Oregon: University of Oregon, 1968, 143-173.

Simsova, S.: *Nicolas Rubakin and Bibliopsychology.* London: Bingley, 1968.

Spang-Hanssen, H.: Kunnskapsorganisasjon, informasjonsgjenfinning, automatisering og språk. I: *Kunnskapsorganisasjon og informasjonsgjenfinning.* Seminar arrangert 3.–7. desember 1973 i samarbeid mellom Norsk hovedkomite for klassifikasjon, Statens Biblioteksskole og Norsk Dokumentasjonsgruppe. Oslo, Riksbibliotekstjenesten, 1974, 7-61.

Taylor, R.S.: Question-negotiation and information seeking in libraries. *College and Research Libraries,* 1968, *29,* 178-194.

Vickery, B.C.: *Information systems.* London: Butterworths, 1973.

9 *Birger Hjørland:* EDB-baseret referenceservice

Edb-teknikken tages i disse år mere og mere i anvendelse som hjælpemiddel for bibliotekernes referencearbejde og dokumentationsvirksomhed. Indenfor forskningsbibliotekssektoren tilbydes der idag tjenester, der bygger på søgning i databaser ved anvendelse af edb. De har eksisteret i en årrække i de teknisk-naturvidenskabelige områder, og er for nylig taget i anvendelse i samfundsvidenskabelige områder som psykologi og pædagogik[1]. Indenfor folkebiblioteksvæsnet er edb-teknikken endnu ikke taget i anvendelse ved litteratur- eller informationssøgning, men på trods af dette spiller edb en vigtig indirekte rolle, idet mange af de referenceinstrumenter, som bibliotekarerne benytter sig af, er fremstillet ved hjælp af edb. Inden vi går over til at se på edb-teknikkens direkte anvendelse i bibliotekernes referencearbejde og dokumentationsvirksomhed, vil vi derfor kaste et blik på dens samlede anvendelsesmuligheder til disse formål, og herunder eksemplificere med nogle anvendelser, der er taget i brug i Danmark.

Anvendelsesområder for edb i bibliotekerne

De anvendelsesområder, der er for edb-teknikken i bibliotekerne, kan inddeles i niveauer, fra relativt simple funktioner, der i princippet svarer til anvendelsen af edb i andre administrative rutiner (f.eks. reservationssystemer til fly, tog og charterrejser), over mere sammensatte funktioner såsom anvendelsen af edb-mediet til opbevaring og afsøgning af f.eks. bibliografiske data, til meget komplicerede som an-

vendelsen af edb til automatiske indexeringssystemer eller automatiske fremstillinger af referater, oversættelser m.v.

Vi vil se på fem niveauer for automatisering[2]:

1 Automatisering af accessionsrutiner, udlånskontrol o. lign.
2 Edb-fremstilling af kataloger.
3 Edb-søgning i lagre (databaser), der er fremstillet manuelt.
4 Fuldautomatisk fremstilling af indexer, referater o.lign.
5 Computer-assisteret fremstilling af indexer, referater o. lign.

Det kan umiddelbart se mærkeligt ud, at den fuldautomatiske fremstilling af indexer m.v. er placeret før den computerassisterede. Dette hænger sammen med, at de fuldautomatiske indexer og referater er temmelig primitive, og at fremstillingen ikke kan automatiseres, hvis produktet skal være bedst muligt. Nogle processer kan automatiseres, andre kan kun udføres af veluddannede personer. Den mest avancerede anvendelse af edb-teknikken får man derfor, når man lader teknikken være en støtte for mennesket.

Det *første niveau* var automatisering af accessionsrutiner, udlånskontrol o.lign. Automatisk udlånskontrol kan f.eks. foregå ved, at de udlånte bøgers nøjagtige identifikation (f.eks. et bognummer i stregkode) aflæses af en optisk læser idet låneren passerer udlånsskranken, og kobles sammen med en entydig identificering af låneren, f. eks. et lånerkort med CPR-nummeret i stregkode, der på samme måde aflæses optisk. Teknikken minder om det system, som visse kædeforretninger har indført, hvor varens data (herunder prisen) står med streger, der kan aflæses med en lyspen, og idet kunden passerer kasseapparatet registreres købet, og edb-maskinen regner dels det samlede beløb ud til kunden, dels gives der besked til lageret om hvor meget der er solgt af de enkelte varer. Biblioteksystemet er dog mere kompliceret ved, at bøgerne skal afleveres igen. Det automatiske system foretager til dette brug udskrift af hjemkaldelseskort på de rigtige tidspunkter. Et andet væsentligt problem i

denne forbindelse er reserveringer, herunder interurbanlån. Der er lavet forsøg med automatisk udlånskontrol både i folke- og forskningsbibliotekerne, f.eks. på Det nordjyske Landsbibliotek, Ølstykke Bibliotek og Danmarks tekniske Bibliotek[3].

Det *andet niveau* var edb-fremstilling af kataloger (herunder fælleskataloger) i bogform udfra manuelt tilvejebragt bibliografisk input og beregnet til manuel søgning. Kendte eksempler er Danmarks Tekniske Biblioteks og Roskilde Universitetsbiblioteks kataloger. Et eksempel på en fagligt afgrænset fælleskatalog, der fremstilles ved hjælp af edb, er den af Rigsbibliotekarembedet udgivne Psykologisk litteratur i danske forskningsbiblioteker. Et meget stort antal af de håndbøger og bibliografier, der anvendes på bibliotekerne, fremstilles nu på denne måde. Der er meget ofte tale om, at udgiverne af sådanne kataloger anvender et lokalt udviklet format, hvad der på længere sigt kan besværliggøre brugen af nationale eller internationale bibliografiske systemer baseret på standardformater som MARC (se side 197). Mere om formatproblemer senere.

Det *tredie niveau* var edb-søgning i databaser, som er tilvejebragt ved en manuel indsamling af data. De fleste af de magnetbåndtjenester, der er operative, hører til dette niveau, og når vi taler om edb-baseret referenceservice, er det anvendelsen af edb-teknikken på dette niveau, vi tænker på. I de fleste tilfælde er der tale om, at man er gået over til at fremstille bibliografier eller håndbøger v.h.a. edb-teknik og fotosætning. Herudfra kan man ved ret få og automatiske omformninger få et brugbart magnetbånd, der kan anvendes til off-line søgning (dvs. søgning uden direkte interaktion mellem søgeren og databasen). Som eksempel kan nævnes en af psykologiens vigtigste bibliografier, Psychological Abstracts, der er grundlagt i 1927. I 1967 gik man over til at fremstille den v.h.a. edb. Resultatet var bl.a. en smukkere lay-out og en hurtigere og nemmere fremstilling af kumulerede registre. Der var ikke økonomisk gevinst ved overgangen til edb, men man forestillede sig,

at man i fremtiden kunne lave mange forskellige bibliografier, når oplysningerne først fandtes i maskinlæsbar form, og at man herved kunne udvide servicen langt mere økonomisk. Magnetbåndsversionen af Psychological Abstracts sælges til informations- og dokumentationscentraler, f.eks. til Karolinska Institutet i Stockholm, der foretager søgninger for brugere i både Sverige og Danmark. Hvordan sådanne edb-baserede søgninger foretages – og hvad der er fordelene ved dem – skal vi se nærmere på i et senere afsnit.

Det *fjerde niveau* var fuldautomatisk fremstilling af registre, referater m.v. (hvad der i fagsproget med en fællesbetegnelse kaldes dokumentrepræsentationer). Et eksempel på edb-fremstillede registre er de såkaldte KWIC-indexer (Key-Word-In-Context, se side 193), der som regel fremstilles udfra dokumenttitler. På fig. 1 er vist et eksempel på en sådan »index«. Denne type index anvendes f.eks. i Dissertation Abstracts International. Der findes andre, lignende indextyper, såsom KWOC (Key-Word-Out-of-Context), hvor nøgleordet er taget ud af sin naturlige sammenhæng for at lette overskueligheden. En helt anden type fuldautomatiske registre er citationsindexerne, der henviser fra et dokuments litteraturhenvisninger til selve dokumentet. Et eksempel herpå er Science Citation Index.

Amerikaneren Gerald Salton har konstrueret et system (»SMART«), der bygger på næsten 100 % automatisk behandling af såvel dokumenttekster som lånespørgsmål. Han mener, at dette er et skridt på vejen mod det dynamiske bibliotek, idet klassifikationen af værkerne aldrig forældes, men hele tiden ajourføres. Han kan godt se begrænsninger i sit system, men peger på, at det er ligeså effektivt som langt mere kostbare systemer, der bygger på menneskelige indexører og litteratursøgere. Problemet ved fuldautomatisk dokument- og spørgsmålsbehandling ligger i, at ord ikke betyder det samme for forskellige mennesker eller for det

KWIC INDEX OF ABSTRACT AND INDEX TITLES

ACOUSTICS abstracts.
AEROSPACE medicine and biology.
International AEROSPACE abstracts.
STAR: scientific and technical AEROSPACE reports.
Adult development and AGING abstracts.
(AGING *see also* GERONTOLOGY)
Journal of studies on ALCOHOL: part B
Abstracts in ANTHROPOLOGY.
BEHAVIOR and physiology index (physiological psychology).

Fig. 1. Eksempel på KWIC-indexer. Fra: Elliott, C. K.: A Guide to the Documentation of Psychology. – London : Clive Bingley, 1971. – p. 105.

samme menneske fra dag til dag. Man kan ikke altid sige, hvad en bog handler om blot ved at trække nogle ord ud af dens tekst (og naturligvis langt mindre ved blot at trække ord ud af titler). Ved en automatisk behandling blot af teksters sproglige udtryk, risikerer man i høj grad netop at miste det nye (og informerende) i tekstens indhold. Det er nøjagtigt de samme problemer, der idag bevirker, at automatisk oversættelse fra ét sprog til et andet ikke kan foretages tilfredsstillende.

Det *femte (og sidste) niveau* var anvendelsen af edb til datamaskinstøttet indexering (og anden dokumentrepræsentation). Dette foregår ved at dokumenter automatisk analyseres for hyppige fagord, der derefter manuelt vurderes og »oversættes« til et klassifikationssystems deskriptorer (emneord). Der kan også være tale om, at mennesket griber ind i en automatisk indexeringsproces og således korrigerer denne. (Dette er således – til en vis grad – muligt i det føromtalte SMART-system). Der er en sammenhæng mellem

de omtalte fem niveauer og indførelsen af edb-teknikken i Danmark, idet simple rutiner naturligvis som oftest bliver indført før komplicerede. Imidlertid viser de store fordele sig først, når man kan integrere funktionerne, idet man hverken kan lave accession eller udlånskontrol før man har en nogenlunde brugbar database. De vigtigste planer for et integreret edb-system til biblioteksbrug i Danmark er FAUST (Folkebibliotekernes AUtomationsSystem), der efter planen er tænkt at skulle omfatte faserne A = accession og katalogisering, B = udlån og C = søgning. FAUST-projektet vil altså omfatte de første tre niveauer, men derimod ikke fuldautomatisk emneindexering. (Udviklingen går snarere i retning af at anvende et system »PRECIS«, der anvendes i den engelske bogfortegnelse, og som forudsætter menneskelig emneordsvalg, men udnytter edb-apparatet til produktion af et omfattende henvisningsapparat).

Databaserne betragtet efter deres indhold

Hvilke oplysninger findes i databaserne? Referencebibliotekarer er vant til at skelne mellem håndbøger og bibliografier og mellem undergrupper som vejvisere, statistiske opslagsværker, sociologiske håndbøger, nationalbibliografier, fagbibliografier og kataloger. Idag spiller de fagbibliografiske databaser den største rolle i praksis, men der kan nævnes eksempler på mange andre typer:

Dokumentsøgesystemer (svarende til bibliografier):
- Nationalbibliografiske databaser (f.eks. British National Bibliography, BNB)
- Fagbibliografiske databaser (f.eks. Educational Resources Information Center, ERIC, og den medicinske MEDLARS/MEDLINE)
- Maskinlæsbare kataloger (f.eks. The National Union Catalog, NU CAT)

Fact-søgesystemer (svarende til håndbøger, benævnes ofte databanker):

Statistiske oplysninger (f.eks. Predicast).
Videnskabelige undersøgelser som spørgeskema-undersøgelser, medicinske data fra patient-journaler eller psykologiske tests.
Vejvisere (f.eks. EIS Industrial Plants; KTAS telefon-bøger hører også til denne kategori, men er ikke offentligt tilgængelige for edb-søgning).
Mødekalendere (f.eks. World Meetings)
Projektkataloger (f.eks. Smithsonian Science Information Exchange, SSIE).

De nævnte eksempler skal blot tjene til at illustrere spændvidden i de eksisterende databaser set udfra et referencetypologisk synspunkt. Vi skal ikke omtale de enkelte databaser nærmere. Søgning i og funktion af de bibliografiske systemer skal vi vende tilbage til, og iøvrigt vil vi her blot slutte med at vise hvordan ét af systemerne, Predicast, på en avanceret måde udnytter datateknikkens fordele.

Predicast er et informationssystem, der skal tjene til planlægningsformål for bl. a. industriforetagender og til dette formål rummer både facts og henvisninger til litteratur. Man kan ud af systemet trække oplysninger om f.eks. befolkningsvækst, indkomst og geografisk fordeling af industriel aktivitet. Systemet tillader en hurtigere og nemmere søgning end hvis man skulle søge efter tilsvarende i trykte medier, men adskiller sig iøvrigt ikke væsentligt i denne henseende. Edb-mediets overlegenhed viser sig ved, at brugeren kan indkode sine egne data, og kan få foretaget aritmetiske og statistiske rutiner for at analysere eller forudsige tidsløb (f.eks. lave prognoser over et bestemt produkts salg). Dette system er operationelt idag, og kan benyttes hvorsomhelst man har en edb-terminal og er tilsluttet Lockheed-systemet i U.S.A. (hvilket man f.eks. har på Danmarks Biblioteksskole).

Maskinlæsbare registreringer

Det man i bibliografi betegner en indførelse, kaldes i edb-sproget en *post*. En post kan også bestå af andet end litteraturhenvisninger, f.eks. statistiske data. En samling af poster betegnes en *fil*. Når en post skal læses af en maskine, er formatet og præcisionen afgørende. Ethvert element (f.eks. forbogstaverne i et forfatternavn) må stå med de korrekte karakterer (dvs. bogstaver, tal, tegn eller mellemrum) på den korrekte position. At det drejer sig om forbogstaverne i et forfatternavn må være entydigt bestemt enten ved den position i posten som dette element indtager eller ved en særlig vejviser (kodefelt), der siger, at det element, der står på den og den bestemte plads i posten er forbogstaverne i et forfatternavn.

En maskinlæsbar post kan ikke skimmes ved et øjekast. I alt væsentligt er det en lineær streng af karakterer, der bliver læst lineært. Enhver karakter bliver repræsenteret af en kode af binære tegn[4]. Den amerikanske standardkode for udveksling af information (ASCII) og den internationale standardorganisation (ISO) har således standarder for, hvorledes karakterer repræsenteres af binære tegn. Disse standarder stemmer ikke ganske overens, men er dog identiske for de væsentligste karakterer. Således skrives et M = 1001101 og et mellemrum eller blanktegn skrives SP = 0100000. Det er naturligvis nødvendigt for edb-maskinen på en eller anden måde at få information om, hvorvidt et bestemt element i en post er en del af titlen, af referatet, forfatternavnet eller andet. Dette problem kan løses på to forskellige måder. Det simpleste er at dele den samlede post op i sektioner af en ganske bestemt længde, således at der f.eks. er afsat 30 karakterer til titlen. Man ved så, at titlen altid begynder et ganske bestemt sted (f.eks. ved karakter nr. 87) og optager pladsen frem til karakter nr. 117. Problemet ved denne form for fikseret feltlængde i posterne er, at der enten må afsættes plads til den længst mulige titel eller at lange titler må forkortes. I mange tilfælde er den

plads, der går til spilde ved titler, der er kortere end det beregnede, meget kostbar. Ved de mere professionelle bibliografiske edb-systemer anvender man derfor et andet system, variable feltlængder, hvor f.eks. titelfeltet, forfatterfeltet, referatfeltet m.v varierer i længde efter behov. For at holde styr på disse variable felter benytter man det førnævnte kodefelt, der registrerer, at i denne specielle post starter titlen med f.eks. karakter nr. 87. For hvert variabelt felt findes en feltkode (f.eks. kan 245 stå for titelfelt). I kodefeltet får man oplysning om det variable felts feltkode, antallet af karakterer i feltet samt startkarakterens position i posten. Ligesom der fandtes forskellige standarder for, hvordan karakterer omsættes til binære koder, er der forskellige standarder for feltkoderne. I MARC-formatet (der især anvendes til bøger) står 350 eksempelvis for bøgernes pris, mens 350 i COSATI formatet (der især anvendes til rapport- og tidsskriftlitteratur) står for emnekoden.

Vi har nu meget kort antydet, hvad det er for principper, der ligger til grund for ordningen af data på maskinlæsbare medier som f.eks. magnetbånd. Formålet har været at give et indtryk af kompleksiteten samt en elementær forståelse for de problemer, der gør sig gældende ved edb-formater såsom MARC-formatet. Det er umiddelbart klart, at standardiseringsproblemerne er mangefold større end ved grafiske registreringer. Foruden de nævnte problemer kommer yderligere, at posterne kan være arrangeret i filerne på forskellige måder, hvad der har stor betydning for søgningens økonomi. Også magnetbåndene kan være forskellige, f.eks. have forskellige sporantal. Standardiseringsproblemer af de her nævnte arter kan i princippet alle løses ved særlige omformningsprogrammer, men der er knyttet stor økonomisk betydning til standardiseringen. Hvad intet omformningsprogram kan klare, er at fremskaffe data, hvis de ikke findes på maskinlæsbar form.

Almindeligvis fremstilles databaserne ved, at dokumenterne analyseres og beskrivende data nedfældes på kodeart, der herefter enten hulles på hulkort eller via en skriveter-

minal overføres til et magnetbånd. Dette magnetbånd anvendes ved produktion af de trykte referenceværker og kopieres og sælges til dokumentationscentraler m.v. Man kan derfor som hovedregel gå ud fra, at de oplysninger, der findes i de trykte udgaver også forefindes på magnetbåndet. Imidlertid kan de ikke altid udnyttes i søgningen, som vi skal se i næste afsnit. Det er ikke altid man anvender alle de data, der findes på magnetbåndet, når man laver trykte værker. Ofte vil man spare papir ved f.eks. kun at trykke de vigtigste emneord i bibliografier til manuel søgning. Den trykte udgave viser derfor ikke nødvendigvis hvilke søgemuligheder der gemmer sig i magnetbåndsudgaverne.

Magnetbånd produceres af stort set samme slags producenter, der producerer omfattende referenceværker: Faglige foreninger, universiteter og forskningsinstitutioner, offentlige organer, private firmaer m.v. Prisen varierer stærkt efter hvem producenten er, således at offentligt producerede databaser ofte er meget billige at købe og benytte. Der skal som regel et stort brugerunderlag til at dække omkostningerne. Båndversionen af Psychological Abstracts koster således ca. 25.000 kr. i årligt abonnement (og købes p.t. i Sverige, men ikke de øvrige nordiske lande). Udgiverorganisationen har copyright til de på båndet værende data, og sælger det mod en skriftlig kontrakt, hvori modtageren bl.a. forpligter sig til ikke at kopiere båndet eller fremstille masseoplag af de litteraturlister, der udarbejdes.

Søgefaciliteter – dokumentationscentraler

Hvis man vil foretage edb-baseret informationssøgning, må man foruden de maskinlæsbare databaser naturligvis være i besiddelse af det nødvendige edb-apparatur. Brugeren må derfor henvende sig til en dokumentationscentral, der køber de pågældende databaser. Her i Danmark produceres så godt som ikke databaser, men der købes en del magnetbånd i udlandet, som så afsøges på dansk udstyr. F.eks. admini-

strerer dokumentationsafdelingen på Danmarks tekniske Bibliotek en del magnetbånd, der køres på I/S Datacentralen af 1959, der har udviklet et søgeprogram kaldet »Teletext«. En stor del af databaseudnyttelsen i Danmark sker dog ikke ved dansk administration af magnetbåndene, men via udenlandske dokumentationscentraler. Således foretager Universitetsbibliotekets 2. afdeling søgninger i MEDLARS/MEDLINE og Psychological Abstracts ved at sende søgeprofiler til Karolinska Institutet i Stockholm, hvorfra man så modtager litteraturudskrifter[5].

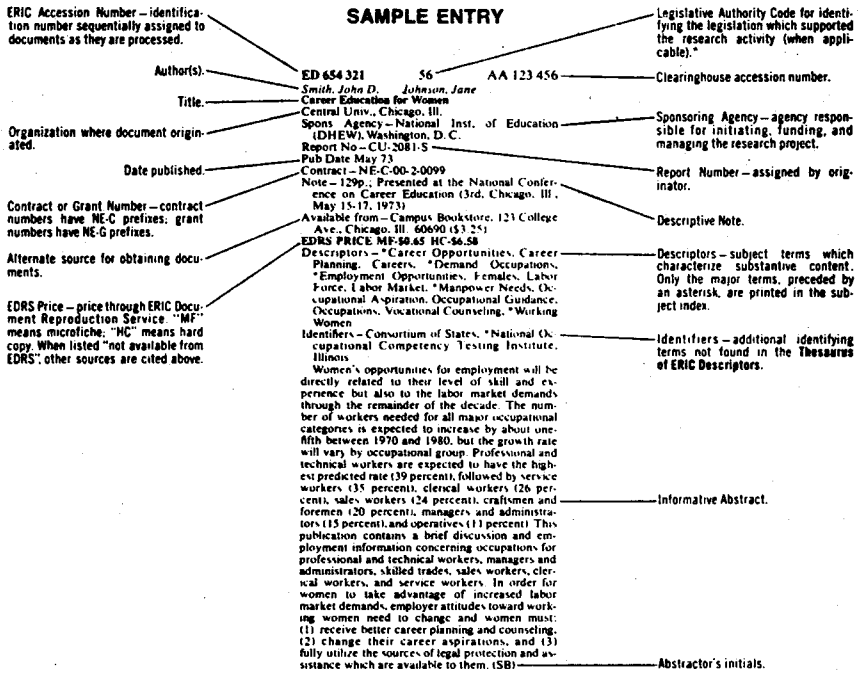
De forskellige datacentraler behandler magnetbåndene vidt forskelligt, de anvender forskellige søgeprogrammer. Dette har stor praktisk betydning for brugeren, idet såvel prisen som kvaliteten af søgningen varierer stærkt efter søgemetoden. Som eksempel kan nævnes, at ERIC-båndene kan søges på Kungliga Tekniska Högskolans bibliotek i Stockholm. Her foretager man en såkaldt *batch*-kørsel, dvs. man samler en mængde søgeprofiler sammen og koder dem ind i systemet. Med jævne mellemrum gennemsøges magnetbåndet (enten de nytilkomne eller det kumulerede), og der produceres litteraturlister, som svarer til de enkelte søgeprofiler. Disse lister sendes til brugeren med post (der er tale om *off-line*). Hvis man søger på ældre årgange af båndene, taler man om retrospektive søgninger. Hvis der kun søges på nye bånd, er der tale om løbende overvågning, *current awareness*. Hvad der f.eks. fra brugerens synspunkt begrænser værdien af søgning fra KTH er, at der ved retrospektive søgninger ikke søges på alle båndene tilbage til 1966, men kun på de 3–4 seneste årgange. En anden væsentlig begrænsning er, at man af økonomiske grunde har udeladt referaterne fra søgningen, således at man hverken kan søge på ord i referaterne eller få disse skrevet med ud i litteraturlisten. Hvis vi sammenligner denne tjeneste med den, som det amerikanske firma Lockheed tilbyder, vil vi se, hvor store forskelle der er. Lockheed tilbyder *on-line*-søgninger, hvilket vil sige, at brugeren via sin edb-terminal kalder systemet, stiller et spørgsmål og får svar på skærmen

eller papirrullen (der er dog også mulighed for at få det skrevet ud på Lockheeds skriveterminaler og sendt med luftpost; denne løsning er fordelagtig ved længere litteraturlister). I Lockheeds system er alle årgange af ERIC-basen tilgængelige (men brugeren kan lade være med at benytte dem alle), og der er referater, som man kan søge på og få med i sin litteraturliste. Det er naturligvis dyrere at få referatet end blot de bibliografiske oplysninger, men kunden har seks valgmuligheder, der spænder fra, at man i det billigste format blot får et accessionsnummer til at man i det dyreste format får udskrevet den fulde post, inklusive referatet.

Når man skal foretage batch-kørsler, sker det ved, at man gennemsøger et magnetbånd med de søgeprofiler, der indgår i computeren. Søgeprofilerne udformes af dokumentalisten, og de overføres herefter til lageret via hulkort eller skriveterminal. Ved on-line søgning er magnetbåndet altfor langsomt et medium og indholdet fra magnetbåndet overføres derfor til et pladelager. Det er imidlertid meget kostbart at have store datamængder liggende på pladelager klar til umiddelbar adgang, og derfor findes et system som ERIC kun på pladelager i et par amerikanske datacentraler. Man kan få adgang hertil fra den øvrige verden via normal telefonforbindelse. De faste omkostninger er relativt beskedne. Fra omkring 15.000 kr. kan man erhverve en anvendelig terminal. Driftsudgifterne består dels af telefonregning, dels af afgifter til henholdsvis databaseproducent og dokumentationscentral. Forfatteren har foretaget en søgning for ca. 20 $ + måske 50 kr. i telefonafgift, men prisen afhænger alene af tilslutningstiden.

Søgestrategier ved edb-baseret søgning

Alle de elementer der indgår i en bibliografisk post (se fig. side 201) – accessionsnummer, forfatternavn, titel, publiceringsdato, udgiverorganisation, emnekoder, referat m.v. –



er i princippet tilgængelige for edb-søgning; men som vi så i foregående afsnit kan der i praksis være tale om, at f.eks. referatet eller andre elementer ikke er søgbare. De elementer, der spiller den helt overvejende rolle for søgningen er emnekoder samt titel og referat. Alle søgbare elementer kan indbyrdes kombineres. Man kan f.eks. søge efter artikler af bestemte forfattere, om bestemte emner, indeholdende bestemte ord i titlen, publiceret et bestemt år osv. Søgning efter ord (eller mere korrekt: tekststreng) i titel eller referat kaldes *fri-tekst-søgning*. Man kan søge de dokumenter, hvis titler og/eller referater indeholder et ganske bestemt ord (f.eks. »automation«). Man kan også søge de dokumenter, der indeholder kombinationer af ord, f.eks. både

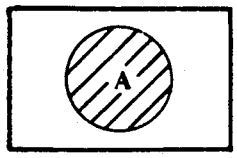
indeholder ordet »bibliotek« og »automation«. Man kan også gå den anden vej og søge på ordstammer som »automat.« Dette kaldes for *trunkering*. Så får man f.eks. fat i alle dokumenter, der indeholder ord som »automation« eller »automatisk« i titel eller referat. Det kan være farligt at søge på for korte ordstammer eller på for almindelige ord, så får man lange litteraturlister fulde af »støj«, og det er meget uoverskueligt.

Foruden fritekstsøgning kan man som sagt søge på emneord. I princippet kan man søge på mange slags emnekoder, f.eks. DK5 tal, men i praksis er det den såkaldt *deskriptorbaserede teknik*, der dominerer, fordi den tillader den mest specifikke emnesøgning og tillader den største fleksibilitet i kombinationen af emner. Vi skal nedenfor kort skitsere princippet i søgning efter denne metode, idet vi viser eksempler på søgelogikker, der anvendes til at kombinere termer i søgeprofiler. (Se s. 203).

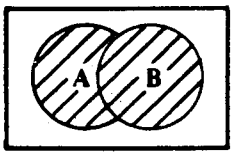
Hvis man f.eks. er interesseret i litteratur om højere uddannelse i Danmark, søger man på deskriptoren »højere uddannelse« og på deskriptoren »Danmark«. De dokumenter, der handler om højere uddannelse i Danmark, er forsynet med begge disse deskriptorer, og man får fat på dokumentet ved at søge på det logiske produkt af disse, altså ved at forlange, at dokumentet skal indeholde begge. Hvis man var interesseret i højere uddannelse i Danmark, men ikke i Grønland, så søgte man som før det logiske produkt af »højere uddannelse« og »Danmark«, men samtidig gav man ordre om, at dokumenter indeholdende ordet »Grønland« ikke måtte skrives ud. Dette kan skrives på følgende måde: »Højere uddannelse« × »Danmark« ÷ »Grønland«. Hvis man foruden Danmark også var interesseret i den højere uddannelse i Sverige, ville søgeprofilen se således ud:

»Højere uddannelse« × (»Danmark« + »Sverige«) ÷ »Grønland«.

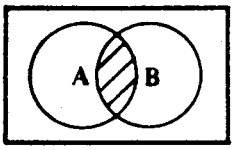
»Danmark« og »Sverige« er to deskriptorer, men hører til samme logiske gruppe. »Højere uddannelse« er en an-



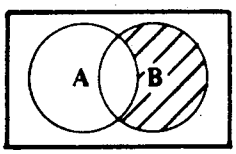
A betyder at referencen skrives ud, dersom der i referencen genfindes en term fra den gruppe termer, der betegnes A.



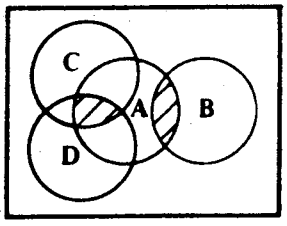
A+B betyder at referencen skrives ud, dersom der i referencen genfindes en term fra gruppen A eller fra gruppen B.



A×B betyder at referencen skrives ud, dersom mindst en term fra gruppen A og en term fra gruppen B samtidigt genfindes i referencen.



B–A A er negeret. Referencen skrives ud dersom den indeholder en term fra gruppe B, men den skrives ikke ud hvis den samtidigt indeholder en term fra gruppe A.



A×(B+C×D) betyder at referencen skrives ud, hvis mindst en term fra A-gruppen og mindst en term fra B-gruppen eller en term fra gruppe A og en term fra hver af grupperne C og D samtidigt genfindes i referencen.

den logisk gruppe, og »Grønland« er en tredie. I edb-baserede søgeprofiler beskriver man blot søgelogikken ved bogstavbetegnelser for de logiske grupper, man opererer med. Disse grupper forbindes så ved operatorerne »og«, »eller« og »ikke«. Ovenstående eksempel kan altså skrives A × B ÷ C. Man specificerer derefter, hvilke deskriptorer de enkelte grupper består af. A = »Højere uddannelse«, B = »Danmark« + »Sverige«, C = »Grønland«. I edb-søgninger er der ofte 20–30 deskriptorer, der på denne måde er kædet sammen i en søgelogik.

Ved edb-søgning skal man være meget omhyggelig med at udvælge emneord. Dette gælder både ved fri-tekst søgning, batch-kørsler og on-line søgning. Ved fritekstsøgning må man være inde i, hvordan fagsproget udtrykkes i titler og referater, mens man ved deskriptorsøgning skal vælge deskriptorer fra en særlig ordliste, kaldet *thesaurus*, der har tre hovedfunktioner:

1 at være en liste over autoriserede indexerings- og søgetermer
2 at udøve kontrol over synonymer og homonymer
3 at henvise benytteren til betydningsbeslægtede termer, overordnede, underordnede og sideordnede begreber.

Ved off-line søgning er valget af de korrekte tekststrenge og deskriptorer særdeles kritisk, fordi man først ser konsekvensen af at anvende en bestemt term, når litteraturlisten foreligger. Ved løbende søgninger kan man ganske vist korrigere sin søgeprofil så ofte man vil, men hvis man f.eks. får én udskrift om måneden, går der en måned før man ser resultatet. Ved retrospektive søgninger off-line bør man sikre sig, at søgeprofilen er tilfredsstillende, f.eks. ved at prøve profilen på en løbende søgning først. I on-line søgning har man mulighed for med det samme at se, hvor mange referencer et søgeord eller en kombination af søgeord giver anledning til, og man kan derfor korrigere sin søgestrategi hen ad vejen. Men også ved on-line søgning er valget af emneord vigtigt, om ikke af anden grund, da fordi det er uøkonomisk at prøve sig frem istedet for at gå frem

Eksempel på søgeprofil fra Kungliga Tekniska Högskolan i Stockholm. Der er tale om løbende overvågning eller SDI (dvs. Selective Dissemination of Information).

ERIC

NAME : SDIZ609 PAGE 01
DATA BASE : ERIC VERSION : 002

RANK : 90 MODIFICATION DATE : 74-04-25
MAX. REFS : 0100 CREATION DATE : 74-02-08

COMMENTS : SELF IDEAL AND SOCIALLY MALADJUSTED

LOGIC : A +B*(C +D) +C*(E +D) +D*E

GROUP	NO	TYPE	WEIGHT	CUM	TERM
A	01	KEY	+02		* SELF CONCEPT TESTS *
A	02	KEY	+02		* IDENTIFICATION TESTS *
B	01	KEY	+02		* IDENTIFICATION (PSYCH *
C	01	KEY	+02		* SELF CONCEPT *
C	02	KEY	+02		* SELF ESTEEM *
C	03	KEY	+02		* SELF EXPRESSION *
C	04	TITL	+02		* SELF IDEAL *
D	01	KEY	+02		* SOCIALLY DEVIANT BEH *
D	02	KEY	+02		* SOCIALLY MALADJUSTED *
D	03	KEY	+02		* EMOTIONAL MALADJUSTMENT *
D	04	KEY	+02		* EMOTIONALLY DISTURBED *
D	05	KEY	+02		* EMOTIONALLY DISTURBED CHILD *
E	01	KEY	+02		* TEACHER INFLUENCE *
E	02	KEY	+02		* PARENT INFLUENCE *
E	03	KEY	+02		* PEER ACCEPTANCE *

TOTAL NO. OF TERMS: 0015

efter en velovervejet strategi. Et andet forhold, man skal være opmærksom på ved edb-baseret søgning, er, at man selv skal kombinere termer for at danne sammensatte begreber. I de konventionelle (med et fagudtryk »prækoordinerede«) indexer findes der jo som regel sammensatte begreber (f.eks. »Danmark, højere uddannelse i«).

Et eksempel på en litteraturliste, der er resultat af en søgning med den foran viste søgeprofil. Referencerne er ordnet efter faldende vægt, en størrelse, der udregnes på baggrund af referencens indhold af deskriptorer. En sådan vægtning er især relevant i off-line søgning, hvor man ikke umiddelbart har mulighed for at korrigere søgeprofilen.

ERIC

74-10-02
ERIC

SDIZ609 7

SOCIAL IDENTITY: PERSPECTIVE AND PROSPECTS 1 M
ZAVALLONI, MARISA
SOCIAL SCIENCE INFORMATION; 12; 3; 65-92 JUN 73
EJ097157

WEIGTH=4.00 *SELF CONCEPT)*IDENTIFICATION (PSYCH*

BASIC GROUP IDENTITY: THE IDOLS OF THE TRIBE 2 M
ISAACS, HAROLD R.
ETHNICITY; 1; 1; 15-41 APR 74
EJ097202

WEIGHT=4.00 *IDENTIFICATION (PSYCH*SELF CONCEPT*

RELIGION, RACISM, AND SELF-IMAGE: THE SIGNIFICANCE OF BELIEFS 3 M
JOHNSON, WAYNE G.
RELIGIOUS EDUCATION; 68; 5; 620-630 SEP-OCT 73
EJ097139

WEIGHT=4.00 *SELF CONCEPT)*IDENTIFICATION (PSYCH*

SOCIAL STATUS OF HEARING IMPAIRED CHILDREN IN REGULAR CLASSROOMS 4 M
KENNEDY, PATRICIA BRUININKS, ROBERT H.
EXCEPTIONAL CHILDREN; 40; 5; 336-42 FEB 74
EJ096152

WEIGHT=4.00 *PEER ACCEPTANCE.*SELF CONCEPTB*

THE MEASUREMENT OF PSYCHOLOGICAL ANDROGYNY 5 M
BEM, SANDRA L.
JOURNAL OF CONSULTING AND CLINICAL PSYCHOLOGY; 42; 2; 155-162 74
EJ095889

WEIGHT=4.00 *IDENTIFICATION (PSYCH*SELF CONCEPTB*

METHODS OF AUTHOR IDENTIFICATION THROUGH STYLISTIC ANALYSIS 6 M
ALLEN, JOHN R.
FRENCH REVIEW; 47; 5; 904-16 APR 74
EJ096288

WEIGHT=2.00 *IDENTIFICATION TESTS&*

EXPLORATORY SELF TESTS 7 M
HARMS, RUTH
ILLINOIS TEACHER OF HOME ECONOMICS; 17; 2; 76-97 NOV-DEC 73
EJ095716

WEIGTH=2.00 *SELF CONCEPT TESTS+*SELF CONCEPT*

Fordele og mangler ved edb-baseret informationssøgning

Hvis man sammenligner manuel og edb-baseret søgning i samme database, kan der opregnes følgende fordele ved edb-søgning:

1 Den tillader *mere komplicerede kombinationer* af deskriptorer og andre søgekoder end manuel søgning.
2 Den tillader søgning på koder, der ikke er tilgængelige for manuel søgning. Her er især fri-tekst søgning væsentlig.
3 Den kan foretages betydeligt hurtigere end manuel søgning. Det er dog en betingelse, at systemet fungerer godt. I nogle tilfælde er der simpelt hen ikke anden udvej end at gennemse alle indførsler i en bibliografi.
4 Den giver mulighed for løbende overvågning, således at brugeren løbende får en individuel liste uden at skulle foretage sig noget aktivt.
5 Den har mange praktiske fordele, såsom større uafhængighed af den fysiske adgang til bibliotekerne og mindre arbejde med at notere referencerne fra bibliografierne.

Af ulemper ved edb-søgning kan nævnes:

1 Den er ret kostbar. Selv når søgningerne er offentligt finansieret, er bevidstheden om prisen ofte en hæmsko for eksperimenter.
2 Selvom on-line teknikken er meget avanceret og i princippet rummer samme muligheder (foruden en del flere) som manuel søgning, så er der velnok knyttet fordele til manuel søgning, der betyder, at denne ikke kan erstattes. Man har ved manuel søgning bedre oversigt over bibliografiens omfang og struktur, man kan bedre skimme. Det såkaldte »serendipity-princip« (dvs. det vigtige forhold, at man finder noget, mens man søger efter noget helt andet) vil ikke træde så stærkt frem ved edb-søgning.

Der knytter sig således både fordele og ulemper til den edb-baserede referenceservice. Det er vigtigt at nå frem

til, at edb-søgning anvendes i de tilfælde, hvor det er mest hensigtsmæssigt, og manuel søgning anvendes til opgaver, hvor dette er mest hensigtsmæssigt. I øjeblikket kan der være en tilbøjelighed til, at de to former for reference- og dokumentationsvirksomhed er administrativt adskilt, således at valget mellem dem ikke er bestemt af egnetheden i den aktuelle situation, men af irrelevante forhold som hvor brugeren tilfældigvis først henvender sig for at få hjælp, hvilke institutioner, der fører gode brugerservicer etc. Referencebibliotekarer lader ofte edb-mulighederne ude af betragtning, ligesom dokumentalister foretager edb-baserede søgninger, uden måske at inddrage det bredere spektrum af manuelle hjælpemidler. Hvis dette problem skal løses, må såvel manuel som edb-baseret søgning være underlagt den samme administrative struktur, og bibliotekarer og dokumentalister må have fuld kendskab til og færdighed i udnyttelsen af begge former, og det må være referencebibliotekaren, der udfra et fagligt kendskab til systemerne træffer afgørelse om valg af manuel eller edb-baseret søgning.

Evaluering af informationssystemer

Anvendelsen af edb til informationssøgning ved anvendelsen af de principper, vi omtalte under søgestrategier (dvs. postkoordinerede indexeringsformer) er først beskrevet af Calvin Mooers, der indførte betegnelsen »information retrieval« (dansk: informationsgenfinding) for denne aktivitet. Calvin Mooers beskrev det ideelle informationsgenfindingssystem som et system, der har evnen til at producere det eksakte sæt dokumenter (hverken mere eller mindre) som brugeren selv ville vælge, hvis han var istand til at læse alle dokumenter i systemet og derpå udvælge de dokumenter, der var relevante for hans spørgsmål[6].

Det var i begyndelsen den almindelige antagelse, at disse edb-systemer fungerede 100 % effektivt og at de totalt over-

flødiggjorde de »gammeldags« klassifikationssystemer i bibliotekerne, men erfaringen viste hurtigt, at der var knyttet visse mangler til systemerne. Først optrådte det problem, at der forekom irrelevante referencer på litteraturlisterne, fordi edb-maskinen slavisk fulgte den givne instruktion og koordinerede termer fra dokumenterne også når disse stod i en sammenhæng, der gav dem en anden betydning. Man taler i denne forbindelse om »false drops«, og dette er et problem den dag idag, især når der er tale om fritekstsøgning. Dette problem har man bl.a. søgt at løse ved at indføre en særlig grammatik (syntaks) i søgesprog, men erfaringerne med sådanne former for syntaks var – i al fald i begyndelsen – ikke gode: systemerne blev meget komplicerede og kostbare, og effektiviteten forbedredes ikke væsentligt. Det modsatte problem opstod også: at relevante dokumenter ikke blev fundet, fordi deres indhold var udtrykt med andre ord eller sætninger end man kunne tage højde for i søgningen. Der meldte sig derfor hurtigt det spørgsmål, hvordan man kunne måle effektiviteten af sådanne systemer, så man f.eks. kunne sammenligne forskellige systemer, forske i hvordan man videre kunne udvikle dem etc.

Man kan tænke sig, at man har et system, der indeholder f.eks. 500 dokumenter. Disse dokumenter kan man gå igennem et for et, og herudfra afgøre, hvorvidt de er relevante i forbindelse med en given problemstilling eller ej. F.eks. kan vi sige, at vi på denne måde finder frem til, at 100 dokumenter er relevante (og hermed at 400 er ikke-relevante). Vi kan nu indexere dokumenterne efter et eller andet system, og foretage en normal søgning f. eks. ved hjælp af deskriptorer. Resultatet af søgningen kan vi f.eks. indsætte i nedenstående 2 × 2 muligheders tabel:

	relevant	ikke-relevant
fremfundet	a	b
ikke fremfundet	c	d

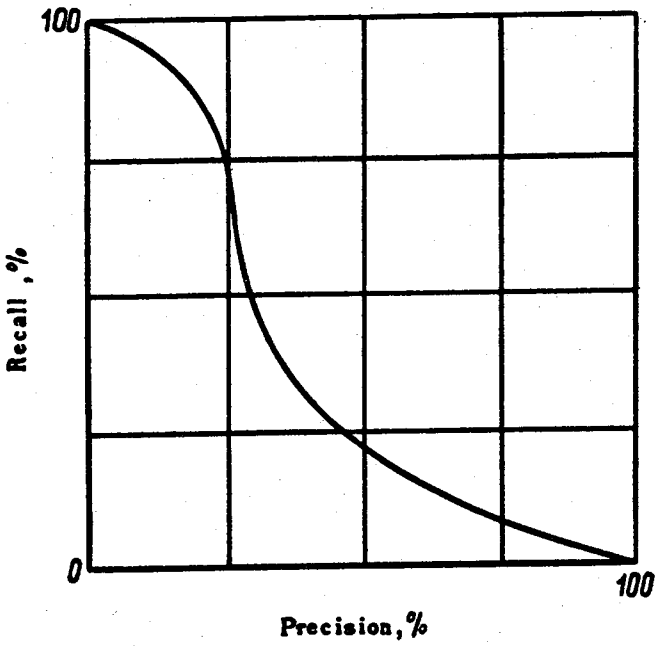
Lad os f.eks. sætte, at en søgning i en katalog over de 500 dokumenter bevirker, at vi får en litteraturliste på ialt 100 referencer, heraf 50 referencer af dem vi på forhånd havde bedømt som relevante. Vi vil så sige, at vi har fundet 50 af 100 relevante dokumenter eller 50 %. Denne størrelse kaldes genfindingsforholdet (»Recall«) og kan altså udtrykkes ved Recall $= \frac{a}{a + c} \times 100\,\%$. Ligeledes kan vi sige, at af de 100 dokumenter vi fandt frem til, var de 50 relevante, dvs. relevansprocenten eller præcisionen af søgningen var 50 %. »Precision« $= \frac{a}{a + b} \times 100\,\%$.

Edb-søgningerne har givet anledning til en del eksperimenteren med anvendelse af Recall og Precision som evalueringsmål. Det vil føre for vidt her at komme nærmere ind på den debat, der hersker på dette felt, men der er almindelig enighed om, at disse mål er særdeles mangelfulde, og at de udførte eksperimenter er tvetydige, så man ikke kan komme med egentlige videnskabelige konklusioner på denne baggrund. Eksperimenterne giver dog visse samstemmende resultater, der kan give et fingerpeg om systemernes effektivitet.

Det må først og fremmest understreges, at enhver tale om, at systemer fungerer med 100 % effektivitet – eller noget der ligner dette – må afvises. Der er meget store forskelle fra system til system og fra søgning til søgning. Der er ikke tale om enkle tekniske årsager til om en søgning er god eller ej, men om et kompleks, hvor spørgsmålets art, litteraturens beskaffenhed, litteratursøgerens kvalifikationer m.v. betyder mere end de rent systemtekniske forhold. Men i almindelighed forekommer der for lav præcision i søgningerne. Visse typiske resultater f.eks. fra den medicinske database MEDLARS viser 70 % recall og 45 % precision som et gennemsnitsresultat af mange søgninger.

Dernæst er det en almindelig iagttagelse, at man i edb-baseret litteratursøgning ofte står overfor et dilemma: Hvis

man vil forøge genfindingsforholdet, så må man søge på flere deskriptorer, men herved får man også mere irrelevant information, således at præcisionen bliver mindre. Man har ligefrem talt om »loven om det inverse forhold mellem recall og precision«, og udfra visse omfattende eksperimenter med edb-søgning fundet frem til nedenstående kurve som et empirisk udtryk for dette dilemma:



Kurven er et udtryk for, at hvis man søger i et bestemt system (f.eks. MEDLARS), så kan man enten anlægge en søgestrategi, der går ud på at maksimere genfindingsforholdet, eller en strategi, der vil maksimere præcisionen. I enkelte søgninger vil man kunne gøre dette uden at det sker på bekostning af den anden variable, men som gen-

nemsnit betragtet vil det medføre, at jo mere man vil maksimere den ene variable, desto dårligere går det for den anden. Det ideelle søgesystem vil da være det, hvor både recall og precision er 100 %, hvilket på figuren vil modsvare et punkt i øverste højre hjørne.

Edb teknikken har på denne måde givet anledning til, at man på en helt ny måde er begyndt at interessere sig for, hvor effektive litteratursøgninger egentligt er. Før edb-søgningerne var der ingen, der forsøgte at måle effektiviteten af litteratursøgninger, f.eks. i folke- eller forskningsbiblioteker. Men i princippet er det naturligvis vigtigt at forsøge at opstille målsætninger for alle former for informationssøgning, såvel manuelle som edb-baserede, og herefter forsøge at evaluere systemerne udfra disse målsætninger. Recall og precision målinger har taget hul på problemet, men som tidligere sagt er der idag enighed om, at disse mål er helt utilstrækkelige, og at de i nogen grad har ført udviklingen i en gal retning. Problemerne har i alt for høj grad været defineret af edb-teknikere, og de tekniske problemer betyder kun lidt i forhold til f.eks. de sprogvidenskabelige problemer, de kognitionspsykologiske problemer, de sociologiske problemer og de videnskabsteoretiske problemer. Den forsker, der først indførte begreberne recall og precision var Allen Kent (der bl.a. er kendt som redaktør af Encyclopedia of Library and Information Science) og han mener, at konsekvensen af, at vi har opbygget informationssystemer udfra nogle fejlagtige mål er katastrofal, og han spår informationssystemerne en hård skæbne[7]: »Men enhver, der hører de ord, jeg har sagt, og ikke handler efter dem, ligner en dåre, som byggede sit hus på sand. Og regnen styrtede ned, og vandstrømmene kom, og vindene blæste og slog imod det hus; og det faldt, og dets fald var stort.« Matthæus 7, 26–27.

Uden at gå så vidt som Allen Kent kan man sige, at edb-teknikken har overforenklet problemerne i forbindelse med informationssøgning, og at der er et stort behov for at forske i informationssystemers ydelser udfra andre og bre-

dere synsvinkler. Der er behov for et skifte i det videnskabelige »paradigme«, der har ligget bag det meste af informationsvidenskaben, dvs. et skifte i det sæt af implicitte antagelser vedr. forskningsmetoder m.v., man har anvendt. Et sådant paradigmeskifte er på vej, og konsekvensen viser sig bl.a. i en større interesse for manuel informationssøgning og for de mere traditionelle klassifikationssystemer, men først og fremmest måske i en mere sociologisk-samfundsvidenskabelig måde at betragte informationssystemer på.

Noter:

1. Erfaringer med edb-søgning i psykologi og pædagogik er nedfældet i *Birger Hjørland & Ole Bredo: Forsøgsvirksomhed med bibliografiske databaser i psykologi og pædagogik.* (ca. 100 s.). Vil blive udsendt fra Danmarks pædagogiske Bibliotek i 1977.

2. Opdelingen i automatiseringsniveauer, samt senere skelnen mellem dokumenters sproglige udtryk og emnemæssige indhold er hentet fra Henning Spang-Hanssen: Kunnskapsorganisasjon, Informationsgjenfinning, automatisering og språk (i: *Kunnskapsorganisjon og informationsgjenfinning.* Seminar arrangert 3.–7. december 1973 i samarbeid mellem Norsk hovedkomité for klassifikasjon, Statens Bibliotekskole og Norsk Dokumentasjonsgruppe. Oslo: Riksbibliotektjenesten, 1974. s. 7–61).

3. Der er udarbejdet forskellige rapporter over danske forsøg med edb-baseret accession og udlån, f. eks.: *Forskningsbibliotekernes edb-projekt for on-line litteraturlokalisering og udlånsstyring.* Rapport. Lyngby: Forskningsbibliotekernes fællesråd, edb-udvalget, 1976. (forsk. paginering).

4. Valget mellem to alternativer, f. eks. 0 og 1 er det mest grundlæggende element i automatisk informationsbehandling. Et sådant element kaldes en »bit« (fra: *bi*nary digit). Alle symboler, f. eks. tal, bogstaver, mellemrum etc. må for at kunne edb-behandles oversættes til sådanne binære symboler. Hvor vi normalt regner i 10-talssystemet, er det for edb-maskinen nødvendigt at regne i et to-talssystem, hvor f. eks. 2 udtryk-

kes ved 10, 3 udtrykkes ved 11, 16 ved 10000 og 32 ved 100000. Den i litteraturlisten anbefalede bog af Mathies & Watson behandler dette emne klart og grundigt.

5. Det skal dog bemærkes, at såvel Universitetsbibliotekets 2. afd. som Karolinska Institutet og Kungliga Tekniska Högskolans Bibliotek nu også anvender on-line søgning, delvis som supplement til batch kørsler. Udviklingen går så stærkt, at det er svært at give en situationsbeskrivelse inden situationen allerede er ændret.

6. Mooers beskrivelse er hentet fra Alan M. Rees: The Evaluation of Retrieval Systems, p. 131. Denne artikel findes i: Arthur W. Elias (ed.): *Key Papers in Information Science.* Washington, D. C.: American Society for Information Science, 1971, pp. 154–169.

7. Allen Kent: Unsolvable Problems. Artiklen findes i: Anthony Debons (ed.): *Information Science – Search for Identity.* New York: Marcel Dekker, 1974, pp. 299–311.

Anbefalet litteratur:

Mathies, M. Lorraine & Peter G. Watson: Computer-Based Reference Service. American Library Association, 1973. 200 s.: ill.
Denne bog er god til at starte på, hvis man vil sætte sig ind i edb-teknikkens anvendelse i bibliotekernes referenceservice. Den er skrevet for at give amerikanske bibliotekarer kendskab til edb-baseret litteratursøgning, og den har valgt et enkelt system *Educational Resources Information Center* som eksempel, fordi det har en bredere interesse end de mere naturvidenskabelige eller tekniske systemer. Systemets opbygning, søgelogik og -strategi gennemgås på en meget klar måde, og disse principper danner grundlag for næsten alle former for edb-baseret referenceservice. I slutningen af bogen kommer forfatterne kort ind på andre systemer af bred almen interesse.

Lancaster, F. W. & E. G. Fayen: Information Retrieval On-line. Los Angeles: Melville, 1973. – 597 s. – (Information Science Series)

En meget grundig og omfattende bog om on-line informationssøgesystemer, hvad der findes, hvordan de fungerer, hvordan man udnytter dem. Forfatteren er især grundig i sin behandling af evalueringsproblematikken.

Keenan, Stella (ed.): *Key Papers on the Use of Computer-Based Bibliographic Services.* Washington: American Society for Information Science, 1973. 179 s.

En antologi med udvalgte engelsksprogede artikler om edb-baserede bibliografiske tjenester.

Annual Review of Information Science and Technology. Washington, American Society for Information Science, 1966–. (årlig).

En årlig oversigt over litteraturen om informationsvidenskab, og de resultater, der er fremkommet. Er særlig grundig på de tekniske områder såsom anvendelsen af edb i bibliotekerne og til referencetjenester. Denne serie kan ikke siges at være velegnet som introduktion, men er nyttig for den, der er inde i grundproblemerne og ønsker et overblik over litteraturen. Gode kumulerede registre gør det let at finde frem til et bestemt system eller problem.

5. ESSENTIALS OF INFORMATION RETRIEVAL

The problem of information retrieval is central to informatics. By now you have learned how scientific data are communicated in the contemporary world, what the sources of scientific information are, what institutions participate in the flow of scientific information and provide for its transfer from the generators to the users and what difficulties arise in conducting information work by conventional means.

Further improvement of this work calls for mechanization and automation of information routines, which in turn requires that these routines be formalized. In order to understand the gist of these processes and to identify their common features, they are embraced by the concept of information retrieval. The information retrieval problem can be viewed as the constant accumulation of an ever-growing bulk of scientific information, on the one hand, and the growth and increasing complexity of information needs of the specialists, on the other hand.

In a generalized form the problem can be seen as the need to select for each user from the totality of the available information only the data he needs at the moment to conduct scientific research or practical work. It is quite evident that no specialist is able to read all existing scientific documents in order to select the essential ones, so they are stored in special collections, and the users wishing to extract from these the necessary documents use a certain procedure which is called information retrieval.

This procedure can be illustrated by a hypothetical example shown in Fig. 16. Let us assume that we have a collection of ten documents. To eliminate the need for scanning the whole file each time the documents containing the required data are to be located, their contents have to be analysed beforehand according to certain characteristics and with a view to possible search questions. If now all the documents are numbered and the characteristics coded by letters, and each document number is linked with its content characteristics, a certain system permitting formalization of the document retrieval process will evolve. Let us assume further that documents represented by combinations of characteristics C and D or characteristics C and F are to be retrieved. Successive scanning of the columns in which the availability of a characteristic is marked with a dot, quickly shows that documents 1, 3, 6 and 8 present the above combinations.

This example shows that to implement information retrieval, it is necessary first to select documents for the system which we call an information retrieval system, and to assign each document a number which will serve as its address in the system. Then a range of characteristics by which these documents will be searched has to be established. This multiplicity of characteristics, expressed - according to certain rules - by terms or conventional symbols (e.g. numbers), constitutes an information retrieval language.

The subject matter of each document is matched against the terms or numbers of this language in order to establish which of the characteristics they represent are contained in the document. The terms or numbers selected as the result of this matching procedure form the search pattern of a document; its formation is called indexing.

During a search, a user's query is also expressed in terms of the information retrieval language and the search request formulation thus obtained is matched against the documents' search patterns. If they completely or partly match, the decision is taken to furnish the appropriate documents to the user. The necessary level of matching is determined by a specified matching criterion.

Basic notions

We shall now define and interpret the basic notions of this section of informatics. Information retrieval is a multiplicity of consecutive operations performed to locate the necessary information or documents containing it, with subsequent retrieval of these documents or their copies, and is effected by means of information retrieval systems.

An information retrieval (IR) system is generally formed by an information retrieval language and a matching criterion designed for information searching in a given information collection. Specific IR systems are realized by means of certain technical facilities, which are nevertheless not included in the abstract notion of an IR system, because the same system can be realized by using different means. These technical facilities, various equipment and machinery will be considered in a separate chapter.

The systems divide into document retrieval and data (fact) retrieval systems. Document retrieval systems produce, in response to a query, documents containing the information sought, their copies, or their addresses in the story. Data retrieval systems are designed to produce facts, e.g. the properties of a particular substance, the structural or molecular formula of a chemical compound, the characteristics of a particular biological species, or the names of those species which possess certain characteristics. The common feature of these systems is that they can retrieve only the information that has been introduced into them beforehand.

At the present time, investigations are under way aimed at creating logical information systems, in which logical processing of information will be possible. Such systems will permit obtaining in answer to a request, not only the previously entered information but also new information which has not been explicitly introduced into the system. Before such systems can be made operational, however, complex problems must be solved, many of which are of a general scientific consequence. The IR systems under review will form part of more elaborate logical information systems.

An information retrieval language is a major component of an IR system. It is a specialized artificial language designed to express the subject-matter of documents and information requests, in order to locate in an information collection those documents which provide answers to certain questions. Every IR language must meet certain requirements. First of all, each notion in it must be expressed by one and only one word (a certain sequence of symbols) and, vice versa, each word must express a single notion. This requirement can be referred to as uniqueness of the vocabulary. Naturally, the IR language vocabulary should not be biassed that is, should not reflect the attitudes of the information generator and user to the given notion. Another requirement, that of a formalized grammar, stipulates that any statement formulated in terms of an IR language must allow a single inter-

pretation. Only when these requirements are satisfied will it be possible to formally match document search patterns and search formulations. It is self-evident that natural languages, in which the same word may have different meanings (homonymy) and the same notion may be expressed in different words (synonymy), where the meaning of a sentence is modified by grammatical means or by context, are unsuitable for information retrieval purposes.

That is why specialized IR languages have long been a standard feature of information retrieval. Conventional kinds of such languages include library and bibliographical classifications (such as those used in a classified catalogue) and alphabetical subject classifications (such as those used in an alphabetical subject catalogue), which we have come across while discussing secondary documents. A feature of these conventional information languages is that all existing knowledge has been distributed in them among the words and phrases in conformity with possible information requests.

However, such a structure is a source of a certain weakness on the part of the conventional IR languages, as the document contents and user requests depend on the progress of science and technology, on the varying interests and personalities of specialists, i.e. to all intents and purpose, they escape accurate accounting and prediction. The setting up of an IR system, the indexing of documents and their retrieval, are operations separated in time, sometimes by very large intervals. For this reason a good portion of information contained in documents that were indexed by means of a conventional IR language will be lost in searching . To obviate this disadvantage, new types of IR languages are created in which the words are not grouped in previously established statements (search patterns). This process takes place either during the indexing stage or even during retrieval. Particular IR languages will be the subject of a large part of the two chapters that follow.

The second component of an IR system, as we have seen, is the matching criterion. A matching criterion is a set of rules according to which the degree of semantic similarity between a document search pattern and a search request formulation is established in an IR system and the decision is taken as to whether to retrieve or not to retrieve a given document in response to a query. Special attention should be paid to the notion of semantic similarity contained in this definition; it has been termed relevance . At the present time this semantic similarity (relevance) cannot be yet established formally and hence automatically. Therefore the relevance judgments, normally made by the user himself, will be necessarily subjective.

The situation is further complicated by the fact that requests are not adequate representations of the specialists' information needs, which are dynamic in character and will continually change with the receipt of new information. When formulating a search request a specialist is not always aware of this actual information need. Getting to know he contents of the documents retrieved may change his idea of the actual need, which will immediately affect his request formulations. Thus, a sort of user-system dialogue takes place, in which the operation of the whole system is optimized through a succession of iterations.

Something like this is observed in everyday life. When a person applies to a consultant, the latter will not supply an answer before he identifies, by posing a number of questions, the actual information need of the applicant. The latter may sometimes gain full realization of his needs only following a talk with the consultant. A strong side of the conventional IR systems, e.g. catalogues of libraries or bibliographic indexes, is that the feed-back to the user is a built-in feature of their structure. Automated document retrieval systems should also be designed in such a way as to enable the searcher to investigate the search file, modifying his search request formulations in conformity with intermediate retrieval results. In other words, they must be designed as systems of the "man-machine" class.

Another way of overcoming this difficulty is the perfection of IR languages and matching criteria. For instance, the matching criterion in a Soviet IR system, named "Pusto-Nepusto" (or "Empty-Non-Empty"), provides retrieval of documents not only when a match of a term or terms occurs in a document search pattern and a search request; if desired by the user, the system can retrieve a document if the terms in its search pattern occupy a higher or lower position in a logical hierarchy than at least one term in a request formulation. Thus, in response to a query concerning documents of which the subject is the properties of liquids, the system can provide documents dealing with the boiling point of water, or vice versa.

Efficiency of an IR system

It has been noted before that a document search pattern recognizes only the main subject treated in the document. Therefore, such a method of document identification cannot ensure the complete recall of all pertinent information in each particular case. On the other hand, part of the retrieved documents may not contain pertinent information at all. These documents constitute the so-called false drops, or noise, and reduce the precision of information searching. The incompleteness and imprecision in information retrieval are a kind of price that has to be paid for facilitation of the search procedure. The price is lower with the more advanced IR languages and matching criteria and, consequently, with the more elaborate searching strategies.

This reasoning brings us to an understanding of the methods of evaluation of document retrieval systems. Operational efficiency of an IR system (economic efficiency will not be discussed here) can be defined as the measure of the system's ability to discharge those functions for which it has been designed. The function of a document retrieval system is to pull out, in answer to a request, relevant documents from an information file, i.e. documents that are semantically related to the query. As a rule, in experiments comparing the efficiency of several IR systems, the documents are assessed for relevance by a team of judges. In practical operation of IR systems, the relevance judgments are made by the user himself, who assesses a document content in terms of his actual information needs. It follows that all evaluation studies are formal in nature and their significance lies mainly with comparisons of different types of IR systems.

The most widely used measures of IR system effectiveness are recall and precision. Recall may be defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents in the information file. Precision refers to the ratio of the number of relevant docu-

ments retrieved to the total number of (both relevant and non-relevant) documents retrieved. Both measures are usually expressed as percentages. The upper part of Fig. 17 illustrates these definitions by a contingency table of relevance and recall.

As demonstrated in the experiments of the British documentalist C. Cleverdon, popularly known as the Cranfield Project, there is an inverse relationship, although not strictly formal, between recall and precision values: an increase in precision will reduce recall. This relationship can be traced in the diagram also contained in Fig. 17. Therefore, increasing the complexity of an IR language and matching criterion, and thereby the searching proceduce generally, with the aim of reducing noise, will result in the loss of many relevant documents. If a 100% precision is sought, i.e. only relevant documents are wanted in the search output, the recall will approximate zero. And, conversely if the aim is a 100% recall, i.e. retrieval of all relevant documents in the collection, the precision will approximate zero on account of the increased noise.

The choice of the recall-precision ratio is of major importance in developing specific IR systems. Most systems have a recall of 70% to 90%, and precision 8% to 20%. These values can be held to be satisfactory for meeting information requests of scientists and engineers, because with a certain information redundancy of the contents of primary scientific documents, the loss of a small proportion of relevant documents will not be felt, and the possibility of iterative searching will help to overcome the adverse effects of noise. When the recall reaches 90%, its further increase will be attended by a sharp drop in precision. It would not be an exaggeration to say that the provision in the search output of the last 10% of relevant documents will demand a higher price than all the previous output. The data cited indicate that there are limits to the efficiency of any IR system, and that both information losses and noise have to be accepted in their operation.

General framework of an IR system

Although the notions of information retrieval and an information retrieval system had their origins several decades ago, the corresponding processes and facilities have been in use in scientific work for hundreds, even thousands of years. It was the achievement of informatics that the operations of a multitude of empirically evolved facilities for information retrieval (reference books, catalogues, bibliographical indexes, libraries etc.) could be fitted into a common framework. The general layout of an IR system is shown in Fig. 18.

We shall analyse its structure in terms of a document retrieval system. You will recall that such a system is designed to retrieve scientific documents in answer to the information requests of the users. Therefore, when an IR system is set up, it should be provided with an initial input of documents, and newly issued documents should be added to it in the future. This process is shown symbolically in the left-hand part of the scheme. The documents come to the Input Converter (IC) where they are subjected to a number of operations. Each document undergoes an analytico-synthetic treatment yielding its search pattern. The search pattern contains a brief formal description of a given document which distinguishes

it from all other documents (bibliographical entry), as well as a concise statement of its subject matter (notation, terms, annotation, abstract).

In addition, each document is assigned an address (usually a number), by which it can be located later. The documents themselves can be converted into a different form (miniaturized or machine-readable). This completes the conversion of the documents at input into the system, and the documents (in the original or converted form) with their storage addresses proceed to the Passive Store (PS). They are kept there until requested by the information users.

The search patterns recorded on some material medium (catalogue cards, punched cards, microfilm, magnetic tape) together with the relevant document addressed, are seperated from the documents and forwarded into the Active Store (AS). There they are arranged in an order facilitating their matching against every incoming request.

The input of information requests into the system is symbolically shown in the right-hand portion of the scheme. First of all, they go into the Input Device (ID) where they are converted from the searcher's natural language into the information retrieval language used in this system. After conversion they emerge as search request formulations, suitable for matching against the search pattern stated in the same information retrieval language. The request formulation proceeds into the Resolver (R) through which are passed the search patterns of the documents from the AS (that is why it is called "active"). There, in the Resolver, the search request formulation is matched against the search patterns. In case of a match (complete or partial, depending on the matching criterion applied), the R gives an instruction to the PS to retrieve the corresponding document by its address. Between the PS and the user is usually an Output Device (OD) which provides for the return of the documents into the PS, and also converts into the original form the documents which have been stored as microreproductions or as machine-readable records. The furnishing of the wanted documents or of their hard copies completes the operation cycle which is started by a user request. Such are in very broad outline the structure and operations of any information retrieval system.

Examples of specific IR systems operation

This can be illustrated using the already familiar example of library operation. The Input Converter (IC) in a library is the cataloguing (processing) department. It is there that a record is made on catalogue cards of the search pattern of a document (bibliographical entry, classification numbers, subject headings) and its address (call number) which is also inscribed on special labels pasted onto the documents (books or journals). The PS is the stacks where the books are shelved according to their call numbers, and the AS is the catalogues, where they are searched. The bibliographers in the bibliographic reference department and the catalogue-room attendants serve as the Input Device (ID), and the reading-room or counter attendants as the Output Device (OD). The Resolver (R) in this case is simulated by the intellectual efforts of the reader who scans the cards in the library catalogue, deducing from the entries which books may be of use to him for solving the problems of his concern.

This example can help to spot some faults of the conventional library as an information retrieval system, in addition to its general shortcomings which were discussed in the previous chapter. They consist primarily in the principles of organization of the Passive and Active Stores.

The PS in the library takes the form of tier stacks, where the document originals are arranged on the shelves. These documents occupy a great amount of space; their issue, shelving arrangement and rearrangement are very labour-consuming and do not lend themselves readily to mechanization. In consequence, the greatest difficulties in the work of modern libraries are those presented by the lack of suitable storage space for the stacks. Besides, the lending of literature can lead to situations where requests for literature in constant demand cannot be met because the books sought are not in the stacks, having been issued to other readers. These difficulties can be overcome only if the documents in the PS are stored as microreproductions or machine-readable records and the readers are issued only true-size hard copies of these documents.

The active storage, realized in the library in the shape of card catalogues, also suffers from serious disadvantages. Because the cards in a catalogue are filed in a linear sequence, any one catalogue can interpret a book collection in one aspect only. We have learned already that the librarians limit themselves to three such aspects: author, subject and classified. Within an aspect a separate card is made for each characteristic for which a search may be conducted.

Thus, increasing the number of aspects and characteristics by which a reader can search his literature will result in more complicated and cumbersome systems of catalogues. This, in turn, increases the labour costs of catalogue maintenance and hampers their use by the readers. For this reason the librarians are obliged to restrict the number of catalogues in libraries and the amount of duplication of bibliographical entries. This, of course, detracts from the search capabilities of the library as an information retrieval system.

This disadvantage of the library catalogues as active stores springs from their inherent organization, which requires that each document must be provided with a search pattern on a seperate information carrier - a catalogue card. A card duplicate with the given search pattern must be reserved for each search characteristic. A card duplication ratio of 1.5 for each catalogue has come to be accepted in library practice. This means that within each aspect a document will have, on the average, one or two search characteristics. But even with this restriction, a library reflecting its stock in three catalogues will have to keep four or five cards per document.

Such an organization of AS is called serial. It presents another inconvenience, namely, that separate searches are necessary for each aspect or characteristic, and multi-aspect searching is difficult. Most mechanized and automated IR systems use an inverted organization of the AS. Essentially it means that a separate information carrier (e.g. a card) is assigned a certain search characteristic or aspect. The addresses of all documents whose search pattern contains this characteristic or aspect are recorded on this carrier.

The inverted scheme has several advantages in comparison with the serial one: it is more compact, it permits multi-aspect searching,and by the feed-back to the user it permits adjusting the search strategy. Adding or deleting certain search characteristics in the request formulation will reduce or expand the search output. These problems will be treated in more detail later on, while discussing descriptor IR systems and technical means of their implementation.

It should be mentioned in conclusion that one important advantage of the conventional library IR systems over mechanized and automated systems is that in the former the functions of the Resolver are performed by the reader himself, who during his searches in catalogues can modify his requests and the intuitively applied matching criterion. This to some extent compensates for the shortcomings of the organization of the AS and makes for higher efficiency of the library IR system as a whole.

In this chapter we have considered the essentials of information retrieval, defined the main concepts of this part of informatics, discussed the methods of assessment of IR system efficiency, the schematic diagram of an IR system and the interactions of its component parts, and the advantages and shortcomings of the different methods of organization of active storage. All this information forms the necessary background for a deeper understanding of the structure and principles of utilization of the existing conventional and automatic IR systems, which will be dealt with in the next chapter.

Questions for self-checking

1. What is information retrieval and what are its basic principles?
2. What component parts constitute an information retrieval system and what are their functions?
3. What is the gist of the relevance problem?
4. How is the efficiency of an IR system determined?
5. What is the purpose and structure of the schematic diagram of an IR system and how do its components interact?
6. Describe the operations of a library and the use of an abstract journal in terms of IR systems?
7. What is the advantage of the inverted organization of AS over the serial one?

Literature

1. Fairthorne,R.A. Towards Information Retrieval. London, Butterworth, 1961, XXIII, 211 p.
2. Kent, A. Textbook on Mechanized Information Retrieval. New York, Wiley, 1962.
3. Sharp, J.R. Some Fundamentals of Information Retrieval. London , Deutsch, 1965, 224 p.
4. Vickery, B.C. On Retrieval System Theory. 2nd ed. London, Butterworth, 1965, XII, 191 p.

6. CONVENTIONAL INFORMATION RETRIEVAL SYSTEMS

Information and bibliographic publications and catalogues of libraries and bibliographical indexes, which were the subject of our third chapter, may serve as particular examples of conventional IR systems or their separate parts.

We revert to them in this chapter in order to consider the principles of their design and operation in terms of information retrieval.

We know already that the basic types of conventional document retrieval systems are the author, alphabetical subject and classified systems. Since in conventional IR systems the matching criterion is not explicitly stated (it is applied by the searcher intuitively), their sole component is the information retrieval language. In the author systems the IR language is formed by the rules for bibliographical description and alphabetization; in the alphabetical subject systems, by lists of subject headings and methods of subject indexing; and in the classified systems, by hierarchical or faceted classifications and classification methods. These conventional-type IR languages will be the subject of this chapter.

Author systems

Author systems originated many centuries ago as bio-bibliographical lists and were designed mainly for the identification, first of different manuscripts, and subsequently of different editions of the same work. The turning-point in their development and use was the mid-19th century when they enjoyed widespread acceptance as author catalogues in large public libraries. It was then that the theory of cataloguing and author catalogue arrangement began to develop giving rise, by the 20th century, to the various national cataloguing codes, and in our day, to the internationally approved principles of descriptive cataloguing.

The main idea behind the rules for bibliographical description is to isolate in every type of work a bibliographic particular (e.g. name(s) of the author(s), title, sponsoring institution) which would best characterize it. These basic data, extracted from a publication according to certain rules and recorded in a symbolic form, constitute what is called an index entry; the totality of these data makes up the vocabulary of an IR language of an author system. The other data, also extracted and recorded according to special rules, constitute the text of the entry and form an integral part of the search pattern of any document in an IR system. The entries are arranged alphabetically by headings in conformity with strictly defined rules which serve as the grammar of the given IR language.

The need for elaborate rules for bibliographical description comes from the IR system requirements which we formulated in the previous chapter, namely, the uniqueness of the vocabulary and the formality of the grammar. But bibliographical data are, unfortunately, far from unique and their manner of presentation varies in different publications. For this reason, descriptive cataloguing methods rigidly regulate the sources, kinds and elements of an entry, as well as its language and spelling. In principle, the source of description is the document as a whole, but primarily its title- or

front-page which is characterized in the language of the original but in modern spelling. Abbreviations of certain words, prescribed by a special list, are used in the text of the entry.

The bibliographical entry contains the following particulars: (1) Author(s), or the heading. (2) Title. (3) Sub-heading data - the data explaining and qualifying the theme and content of the document, its nature and purpose; data concerning the approval or endorsement of official documents; data on the persons participating in the compilation of the document or its preparation for publication; data on the edition or the official character of the publication. (4) Imprint - place of publication, publishers, and year of publication. (5) Collation - number of pages, and the presence of illustrations. (6) Information before the title - the name of the institution sponsoring the publication; the title of the series in which the publication appears, and its serial number. (7) Bibliographical notes - data on the authors given in the publication (if the notes disagree with the data accepted for the entry, or if the publication has more than two authors); data on the variant readings of the title, on the availability of a bibliography in the document, on publication defects. Fig. 19 gives a break-down of bibliographical data on the title-page of a publication and in its entry.

Different approaches to descriptive cataloguing arise in connection with the authorship, number of volumes, frequency of their appearance and the polygraphic independence of a publication. Authorship is the most individualized attribute for some of the more widespread publication types. Works by one, two and three authors are entered under the surname and forename (initials) of the first author. An added entry is made for the second and third co-author (or they are listed in the alphabetical index of names), so as to reflect a given work in the common alphabetical sequence of their works. The names of these authors are presented in the headings in a unified form, in order to ensure the unique location of the works of the same author in the common alphabetical listing.

The authors are called compilers in such publications as dictionaries, guides, handbooks, study aids and guidance manuals, reading books, anthologies, technical and scientific information material, advanced methods exchange material and fundamental bibliographical indexes. All these types of publications are indexed under their compilers, who have done a piece of independant research.

Works by four or more authors, collections of papers of different authors, syllabuses, procedure manuals, statistical handbooks and similar publications are entered under the title, even where the compiler or editor are indicated. (In some countries, e.g. in English-speaking countries, such publications are indexed under compiler or editor). Added entries are made for the first author, if a work of four or more authors is indexed, and for the compiler or editor, if a collection is indexed, which is a great help to those users who will remember a book by such names rather than by its title.

Official publications are issued on behalf of institutions, organizations or agencies as their documents for which they bear full responsibility and which are directly related to their activities. These include statutes, regulations, manifestos, appeals, reports, plans, decrees, resolutions, instructions and other similar material. Their titles usually begin with these standard words and contain the name of an institution or organization,

e.g., "The Charter of the Communist Party of the Soviet Union". The salient feature of an official publication, which determines information requests for it, is the name of the issuing body. For this reason official publications are commonly indexed under the name of the institution, which in this case appears as a corporate author (not to be confused with joint authors, viz., several persons who have collaborated in writing a work). The same indexing procedure is used for many non-official publications, such as "Transactions","Proceedings" or "Papers" of scientific institutions, congresses, symposia etc. if their title begins with or is confined to these standard words, and contains as an integral part the name of the organization or institution; for example, "Transactions of the Krupskaya Institute of Culture in Leningrad".

According to the number of volumes (issues) and their frequency of appearance, publications are divided into multi-volume and series works, and periodicals. The typographic make-up of a multi-volume publication or a series is such that much bibliographical data recurs in each of its volumes or issues. For such publications a series entry is made, with the general part giving all the recurring details, and the specification listing only those details which distinguish one volume from another. This saves a lot of entry space and indexing time. It is important though that not only the similarities but also the differences between multi-volume books and series should be realized. A multi-volume publication is limited by its content to a definite number of volumes, while a series comprises quite independent publications united only by their common nature or by the similarity of their subject matter. Therefore, the series entry for a multi-volume book is its main - and normally its only - entry, giving full bibliographic details. For series publications, on the contrary, main entries will be the entries for individual issues, which ought always to be made. Here the series entry is an auxiliary one; it is very brief and made only in those cases when a series as a whole can be of scientific interest.

Periodical publications, which include serials, journals and newspapers, are also indexed by means of a series entry, usually under the title. The only exception to the general rule is constituted by the serials which occupy an intermediate position between the book series and the journals. Occasionally, they are entered under a corporate author with, in addition, separate entries made for the individual volumes or issues which may have an author or be collections of works having a title of their own.

To index a part of a publication which is in itself an independent work (e.g. a paper in a collection or journal), an analytical entry is made under the author and title, as with other independent publications. In place of the imprint however details of the publication in which the paper is included are indicated. These details may vary according to whether the publication in question is a collection, serial, journal or newspaper. An entirely different matter is making an analytical entry for a part of a work, such as a chapter or section of a book, which cannot be viewed autonomously out of the context of the publication to which it belongs. Therefore, chapters or sections are entered under the author and title of the relevant publication. But instead of the collation such an analytical entry gives the number and title of the chapter or section indexed, as well as the number of its pages in the book.

The same procedure applies to reviews, or critical comments on books, which usually appear in journals, serials or collections of papers. What the reader wants to know is primarily whether there is a review for this or that book (often he will not know who wrote this review), what title it has and where it was published. For this reason reviews are entered under the author and title of the book reviewed. The entry lists all the particulars which are usual in book entries, but instead of collation data it gives details of the review and of the publication in which it appears. The most important reviews may also be indexed as separate papers. The above rules for description of the main types of publications are schematically displayed in Fig. 20, with illustrative examples shown in Fig. 21.

Rules for alphabetical filing of entries, which form the grammar of this conventional IR language, specify the various methods of such filing within one and within several natural languages. There are two commonly used methods of alphabetization of words, names or sentences, which are both based on the order of letters in the first words. By the first method ("word-by-word"), entries are filed in the sequence of words, by the second ("letter-by-letter"), strictly in the sequence of letters of the alphabet, without consideration for the division of sentences into words. These methods can be explained by the following:

"Word-by-word"	"Letter-by-letter"
New England	Newark
New wives for old	New England
Newark	Newman
Newman	New wives for old

Most alphabetical systems use the "word-by-word" method as better common-sense and less formal. It applies not only to prepositions and other function words (articles, particles), but also to complex and compound words, including forenames and surnames. In some systems this rule is made still more complicated: if the first words coincide, it is not the alphabetic sequence of the words that follow which is taken into consideration but the heading category - entries are listed first under the names of the authors, then under the names of institutions, and finally, under the titles beginning with the same word. Since this facilitates locating entries in long alphabetical sequences, the practice of listing these three categories of headings in three different alphabetic sequences is sometimes adopted.

There is another range of problems in alphabetization of index entries connected with the fact that different languages use different alphabets and systems of writing. Most of the nationalities of the USSR, for instance, make use of Russian letters; many nations in Europe, America and Africa use Roman letters; some nations of Asia employ national systems of writing, e.g. Arabic, Ancient Indic (Sanskrit), Chinese hieroglyphics, etc. Moreover, different languages using the same letters may have different alphabets. Thus, the alphabet of the Azerbaijan language differs from that of the Russian language, and the English alphabet differs from the Polish one. Because of this, entries in languages which have different letters and alphabets have to be placed in different alphabetic sequences. If occasion demands a common alphabetic sequence for entries in different languages within the same system of letters, the basis is provided by some artificial alphabet that usually ignores auxiliary letters and diacritics.

In conclusion, it should be noted that until quite recently alphabetic systems have been the only means of identification of publications, of searching for specific titles in IR systems, and of searching for publications by specific personal and corporate authors. The excessive work and time-consuming efforts resulting from the existing cataloguing rules and their insufficient standardization at the international level should be mentioned as the principal faults of those systems. The recent proposal to use standard numbers for books and serials as their identifiers in information work, is an attempt to rectify those inadequacies. The main idea of the proposal is to assign to every new book and every serial title a unique number associated with the respective entry in a national bibliography. Implementation of this proposal will considerably facilitate the input of documents in automated IR systems and the exchange of bibliographical data both on a national and international scale.

Subject systems

We have already seen, as in the case of permuted-title and citation indexes, that a listing of some particulars of a bibliographic entry can, under certain conditions, provide a subject searching capability. This fact has long been utilized in subject (or alphabetical subject) systems. Subject catalogues and alphabetical subject indexes, which were discussed in the second chapter, are examples.

Subject cataloguing is a special kind of document classification which is concerned with the concise formulation of document contents by means of one or several standard words called subject headings. These are arranged alphabetically and followed by either bibliographical entries, as in subject catalogues, or by classification numbers, bibliographic entry numbers or book page numbers, as in alphabetical subject indexes.

The task of subject cataloguing is to identify the subject which is the main theme in a document, as well as to establish its main features and its relationships with other subjects. The different features of a subject are as follows: its history(origins and development), structure, composition, properties, state, purpose, interaction with other subjects, investigation, evaluation, etc. In subject cataloguing, documents are grouped by subjects not belonging exclusively to any one area of knowledge, which makes it possible to form different complexes ("subjects") such as a personality, human activity, geographical feature, natural phenomenon, social phenomenon, material, property, etc. Thus, the subject heading "X-rays" will bring together documents which are concerned with the nature of X-rays(physics), their use in medical diagnostics (medicine), and design of X-ray apparatus (X-ray engineering). As subject headings are listed in a common alphabetical sequence, it is not necessary - while searching for a document on a specific subject - to know the place of this subject in the general systematics of scientific knowledge. Subject headings of quite different content will be listed side by side in alphabetical subject systems, e.g.:

Apricots
Archangelsk
Aristotle
Automobiles.

On the other hand, alphabetical arrangement brings together words of the same root which often prove to be associated semantically; this will help to create in a subject system a broad thematic complex, e.g.:

Atom
Atomic bombs
Atomic electric power station
Atomic energy
Atomic ice-breakers
Atomic nucleus
Atomic spectra
Atomic thermal capacity.

A subject heading may be selected from a special previously established list or it may be formulated by the indexer; in either case it recognizes only the main aspects of a document's content, which in practice is usually covered by a very few headings. The major problems faced in building up the IR language of a subject system are the use of "generic" and "specific" headings, the inverted word order in the formulation of subject headings, and the use of sub-headings. Without going deeply into the methodology of subject classification, solutions for each of these problems can be illustrated by a few examples.

A subject heading must be most specific in defining the subject of a document. Thus, a document dealing with motor-car carburettors should be entered under "Carburettors" and not under "Internal combustion engines". In this case the subject heading is *specific*, or rather, adequate, i.e. corresponding to the main subject of the document. If, in anticipation of possible requests by users of the system, it is thought desirable to enter a given document under a heading expressing a more general idea (in our example, Internal combustion engines), such a heading is considered to be *generic* or, more correctly, generalizing.

In order to bring together related subject material in a subject system, the words in a heading can be regrouped (inversion). Thus, an adjective is often placed after a noun, e.g.:

Microbiology - agricultural
Microbiology - industrial
Microbiology - medical
Microbiology - veterinary.

In some cases, however , inversion is not necessary; for example, when an adjective defines an important attribute of a subject (Computational technology, Electronic valves), or in the names of historical events and other fixed phrases(Roman Empire). In order to divide the material within a subject heading, subheadings are used, e.g. topical(Water - Analysis, Evaporation, Purification, Chlorination), geographical (Animal Breeding - USSR, Great Britain, France, USA), form (Standards - Bibliography, Classification Schedules, Indexes), chronological (Periodicals - History - 17th Century, 18th Century, 19th Century). A special position is occupied by the standard or common subheadings, e.g. those used in the subject cataloguing of the literature on materials (Concrete - Analysis, Defects, Substitutes, Testing).

Subject systems are most effective in searching for narrowly defined subjects and are far less useful for broad thematic searches in a particular area of knowledge. This is their peculiarity, not to be viewed either as an advantage or disadvantage - it merely defines the conditions of their application. On the other hand, subject systems have a number of unquestionable advantages. They are simple in operation and require no preliminary training on the part of the user; introduction of new subject headings does not involve any changes in the existing headings and re-indexing of the documents already in the system. Growth in the number of subject headings and documents in a system does not create any special difficulties either for the indexers or for the users. Subject systems are easily implemented as either manual card catalogues and card files or as bibliographical lists.

But subject systems also suffer from a number of shortcomings. They retrieve only those documents whose main subjects match document requests, and they do not give details on other topics discussed in scientific documents. (This, however, is the common inadequacy of conventional IR systems oriented to manual handling). Another shortcoming which is equally common to all conventional IR systems is the difficulty of multi-aspect searches, since a subject system does not permit simultaneous retrieval of information on several documents entered under different headings, that is to say, it is impossible to conduct a search using any combination of characteristics, or their logical product. The retrieval languages of subject systems employ the vocabulary of a specific natural language and are therefore unsuitable for international use. Finally, in comparison with other conventional IR systems, the setting up and maintenance of a subject system is a very laborious task for highly skilled personnel.

Hierarchical classifications

Classification is the grouping of subjects or relations in classes according to a common characteristic which is inherent in all the subjects of a given kind and distinguishes them from the subjects of the other kinds, the grouping being done in such a way as to put each class in the system in a definite, fixed position with respect to the other classes. The characteristic according to which classification is done is called a principle of division.

The classification process obeys some formal rules of logic: subjects or relations may be divided according to only one principle of division at a time, and the resulting classes and subclasses should be mutually exclusive, the division into classes should be balanced and continuous, without leaps. Classifications, in which every subclass has only one class that immediately precedes it(relations of strong hierarchy) and all subclasses subordinated to only one more general class (collateral subordination relations) are called hierarchical. By way of visual demonstration of a hierarchical classification, Fig. 22 shows its graph, that is a diagram consisting of points (apices) and the lines (edges) that connect them, as well as an Euler-Venn diagram indicating the relationships between the classification divisions.

Classification of documentary information, which is one of the most important types of an IR language in conventional systems, consists in the grouping of scientific documents by areas of knowledge according to their content. Such document classifications are more or less associated with

classifications of sciences, as their main classes will broadly correspond to specific fields of knowledge, and further division of these classes will correspond to the structure of these fields of knowledge.

However, these classifications, which have been called library classifications, are not the same as classifications of sciences. Library classifications must be built according to some formal rules of logic, because only under this condition can they serve as IR languages and provide unique identifications to the documents. But rules of logic are inapplicable to classification of sciences, since there are no clear-cut boundaries between different sciences. The other distinctions also spring from the strictly practical nature of library classifications and the peculiarities of the subjects classified. Apart from the division by document content, a library classification must include division by types of publications (books, periodicals, special types of technical publications), by their purpose (scientific, popular science, instructional), by language, etc.

Library classification schemes are generally published as schedules consisting of two parts: the main tables and the tables of auxiliaries or common subdivisions. Main tables list all fields of knowledge and their sections in a logical order with each level of division according to only one characteristic; each heading may have a number of subordinate headings forming a certain hierarchy. Because of these characteristic features, conventional library schemes are sometimes called linear-hierarchical. Tables of auxiliaries or listings of common subdivisions display the recurring characteristics of different subjects, e.g. the characteristic of place and time to which a document content refers, or that of type of publication.

Each division of a library classification is assigned a conventional symbol called a code number. Classification code numbers constitute the notation which may be numerical, alphabetical, or mixed. The merits of a numerical notation are based on the fact that a numerical sequence is more obvious and familiar than an alphabetic one, that any combinations of digits are easily pronounceable as numbers or figures, and that the Arabic numerals are understood by all nations whatever their spoken or written language.

At the same time numerical notations have a serious drawback, namely their limited base: because there are only ten digits (from 0 to 9), very lengthy numbers with many digits have to be formed in order to designate complex or very specific notions. A solution is frequently sought in the use of mixed code 'numbers' which include letters as well as numerals.

From the standpoint of structure, notations may be non-structural and structural or hierarchical. Non-structural notations make use of the serial numbers of classification subdivisions in a common numerical sequence. Such notations have little mnemonic value, do not reflect the hierarchy of classification subclasses, and make further division of these subclasses difficult. For this reason, most of the modern library classifications employ structural notations which may be either numerical, alphabetic or mixed. Structural notations represent the conceptual structure of a classification, since each class is designated by one symbol, and all primary divisions of this class by two symbols, of which the first stands for the class, and the second for the corresponding subclass. Structural notation permits the detailing of a classification scheme to any desired length or depth.

Universal Decimal Classification

Library classifications have been with us from very early times, and this is hardly surprising, as the need for them arose simultaneously with the emergence of written records. Nowadays, hundreds of different classifications are in use.

The Universal Decimal Classification is certainly the most significant of the classification schemes developed around the turn of this century, and also the most extensively used in all countries of the world. It owes its origin to the work of two eminent Belgian bibliographers, Paul Otlet (1868-1944) and Henri LaFontaine (1854-1943). The international conference of bibliographers , which was convened in 1895 through their efforts, formed two organizations: Bureau International de Bibliographie and Institut International de Bibliographie which were later to unite under the second name (IIB). The Institut International de Bibliographie which had its headquarters in Brussels was entrusted with the task of compiling a Universal Bibliographical Repertory, i.e. a bibliography that was to cover the literature in all fields of knowledge published in all countries in all languages from the earliest times to our day. This obviously unfeasible task was never realized but required for its fulfilment the provision of a depth classification scheme encompassing all parts of knowledge, i.e. a comprehensive or universal classification, and usable on an international scale.

The basis for this new classification scheme, developed by a group of experts headed by P. Otlet, was furnished by the Decimal Classification of Melvil Dewey, which had been considerably revised and extended. The new version of the decimal classification has appeared in French in separate issues since 1897: in 1905 it was issued as a complete publication, and in 1907 reprinted under the title "Manuel du répertoire bibliographique universel".(1) Subsequent editions of this version and other versions in different countries and different languages, have been sponsored by the International Federation for Documentation, successor to the Institut International de Bibliographie,and are known as the Universal Decimal Classification or UDC.

The UDC comprises main tables, auxiliary tables and an alphabetical subject index. The main tables display the numbers by which the documents are systematized according to their contents; each concept presented in the main tables must have a definite UDC number. The sum total of human knowledge is divided into 10 main divisions, shown in Figs. 23 and 24. Each successive digit added to the designation of a main division does not change its general content but merely qualifies it.Such a method of building UDC numbers makes it possible to divide any general idea into narrower specialized topics. The greater the depth of division, the longer will be the notation. Thus, 'compressive strength of soils' in construction engineering is denoted by 624.131.439.4. To facilitate the reading of the notation, every three digits are followed by a full-point. The numbers are pronounced as a sequence of integers, e.g. six-two-four (point) one-three-one (point) etc. Fig. 25 gives an example of interpretation of a UDC number, 621.22 Water power. Hydraulic machines.

1. Manuel du répertoire bibliographique universel. Bruxelles, 1907.

UDC numbers are filed, within each division, according to the principle 'from the general to the specific'.Their order depends only on the sequence of figures in each division and not on the length of a number.

In the main tables are listed, apart from the main classification numbers, special (analytical) subdivisions which represent the characteristics typical of a limited range of concepts; they are limited to use in the section under which they are listed. If applicable in many subdivisions of a given section they are joined to the main numbers by a hyphen; if applicable only in the subdivision concerned, they are joined by means of .0 (point 0).

The auxiliary tables list the general or common auxiliaries, which represent recurring characteristics according to which documents can be divided in all areas of knowledge and which are applicable in all the sections of UDC, as well as special signs which serve to join several main UDC numbers. The common auxiliaries may be of language, of form, of place, of time, of race and nationality, of viewpoint. They are used in all sections of the scheme with the same meaning. When occasion demands, letters of the Roman or Cyrillic alphabets, separate words or names can be attached to UDC numbers,e.g. 92 Einstein, for a biography of Einstein.

For a more exhaustive and accurate representation of the subject-matter of documents, the UDC permits - in addition to the main numbers and auxiliaries - the connection of several numbers with the help of different symbols. The common auxiliaries and the connective symbols are shown in Fig. 24. The addition (plus) sign is used in those cases when a document deals with several separate concepts of equal value. The extension (stroke) sign is used to join the first and last of a series of consecutive numbers denoting adjacent related concepts. The relation (colon) sign permits the classifying of documents dealing with concepts that are conceptually related. The synthesis (apostrophe) sign is used, inter alia, for documents on chemical compounds and alloys.

The task of classifying by UDC is made much easier by the availability of an alphabetical subject index to the tables. Having established the primary subject of a document one must look up in the index the UDC division or subdivision that corresponds to the document content, then scan the array of the relevant section, select the most appropriate heading and write down its number. In establishing the heading it is not necessary to follow the last level of division in the tables used. One should be guided here by the availability of material on this subject and the prospects of its growth. The use of too general numbers is undesirable, however, because the depth of classificatory analysis of a document will then be sacrificed.

In classifying documents in the mathematical natural and applied sciences (UDC divisions 5 and 6), the most important action is to delimit their topics. It should be remembered that the natural sciences study the laws of nature, while the applied sciences use their findings for practical purposes. X-rays, for instance are studied in physics, and the different aspects of their application in technology and medicine: hence, documents describing direct observations or results of laboratory studies are entered under 537.531 X-ray and Gamma-ray physics (non-corpuscular rays in discharges); documents on the design and manufacture of X-ray apparatus under 621.386 X-ray tubes and accessories (electrotechnology) and material

on the diagnostic application of these rays under a subdivision of medicine, 616-073.7. The use of other subdivisions is also possible here: for example, if a document deals with the subject of films for X-ray photography it should be indexed by 771.531.34 in Photography and cinematography.

The UDC has the following specific features: (1) coverage of all fields of knowledge (the recent UDC editions contain more than 100.000 subject headings); (2) the decimal principle of division, which permits unlimited division of subclasses without violating the basic structure of the scheme; (3) the use of exclusively numerical notation, which is relatively easy to memorize and equally comprehensible to specialists using different languages; (4) the availability of an elaborate system of auxiliaries; (5) the use of the principle of synthetic notation; (6) the possibility of classifying any number of documents to any level of division.

The primary distinctive features of the UDC are its elaborate system of auxiliaries and the synthetic notation. It was precisely because of these features that the emergence of the UDC in a way meant a breakthrough in library classification. The former 'enumerative' schemes with previously established headings and ready-made numbers were ousted by a new flexible classification in which the necessary headings are formed during the classification process either by combining numbers with auxiliaries or joining together two or more numbers. The existence of a well organized international system for keeping the UDC up-to-date is its unquestionable merit and places it in an advantageous position with respect to other present-day classifications. Naturally, the maintenance of such a system takes much time and its operation presents great difficulties and entails considerable expenditure, but it is only through such a system that the development and perfection of a classification is possible.

All the same, it should be noted that the division of the whole universe of human knowledge into a mere ten classes, and the order of these classes in the UDC, do not correspond to the present level of the development of science. Despite this and some other defects, the UDC has become widespread and found application in thousands of institutions in many countries of the world. Presently the size of the complete UDC schedules is approximately 10 volumes each of 300-500 pages.

In conclusion, it would not be out of place to list the principal merits of the UDC which are: universality, international usability, the decimal system of notation, the number-building principle and a smoothly functioning body for keeping it up-to-date. At the present stage of development of classification theory and practice, the UDC remains the only internationally accepted universal system capable of revealing the contents of reference collections in sufficient detail, ensuring speedy information retrieval, and making for closer international co-operation. These merits of the UDC point with certainty at the advantage of using it in scientific information agencies and in special libraries for classifying literature in the natural sciences and technology. The Council of Ministers of the USSR decreed on 11 November 1962 the compulsory use of the UDC by information centres and technical libraries throughout the country for classifying literature in the natural and technical sciences.

Decimal classification has been in use in the USSR since 1921. The public libraries still utilize the different versions of the library classification schedules which are based on a modification of UDC made by So-

viet bibliographer L.N. Tropovsky. Work on a new Soviet scheme of classification for research libraries, based on the Marxist-Leninist classification of sciences, has been going on for many years; it has been appearing in separate issues since 1960 (1).

The new Soviet scheme makes wide use of the techniques which determined the success of the decimal classification: the decimal structure of notation for the subdivisions of each class, the synthetic principle, the minutely elaborated tables of common subdivisions. All the same, the mixed (numerical-letter) notation employing different punctuation marks seems rather complicated in comparison with the UDC. Below are the main divisions of the scheme:

А Marxism-Leninism
Б Natural sciences in general
В Physico-mathematical sciences
Г Chemical sciences
Д Earth sciences (geodesical, geophysical, geological and geographical sciences)
Е Biological sciences
Ж/О Engineering and technological sciences
П Agriculture and forestry. Agricultural sciences
Р Health protection and medical sciences
С Social sciences in general
Т History. Historical sciences
У Economics. Economical sciences
Ф Communist and workers' parties. Socio-political organizations of labour
Х Government and law Juridical sciences
Ч Military science
Ц Culture. Science. Education
Ш Philological sciences. Literature
Щ Art. Art criticism
Э Religion. Atheism
Ю Philosophical sciences. Psychology
Я Literature of universal content.

Specific features of hierarchical classifications

Classification is one of the major tools of humanity in the process of cognition, one of the normal methods people use to define an object. V.I. Lenin wrote in this connection: "What is meant by giving a 'definition'? It means essentially to bring a given concept within a more comprehensive concept". (2) This is one of the strong sides of hierarchical classification as an information retrieval language.

1. Bibliotechno-bibliograficheskaya klassifikatsiya. Tablitsy dlya nauchnykh bibliotek. Vyp. I-25. M., 1960-1969. (Gos. b-ka im. V.I. Lenina). (Bibliothecal-Bibliographical Classification. Schedules for research libraries. Issues 1-25. Moscow, 1960-1969. (Lenin State Library of the USSR)).

2. V.I. Lenin. Collected works. Vol. 14. Moscow, Foreign Languages Publishing House, 1962, p. 146

However, it should be recognized that hierarchical classifications, like any other IR languages, have certain limitations. It will be recalled that the development of science shows two conflicting tendencies, namely differentiation and integration, or the division of scientific fields into ever new trends and disciplines, on the one hand, and the interpenetration of the related and even remote fields and disciplines, on the other. The process of science differentiation is more or less adequately recognized in the linear hierarchical classifications, but the integration and interpenetration of sciences cannot be adequately reflected in these classifications. For example, it is an extremely complex task to find in the UDC a place for the scientific trends and topics which emerge on the borderlines of chemistry, geology and biology, or of mathematics and linguistics.

This limitation of hierarchical classifications, which strictly delimits separate sciences in accordance with formal rules of logic, runs counter to the synthesizing trend in the progress of science. The same rules will not permit multi-aspect indexing of documents by means of an hierarchical classification, and information searching for any combination of characteristics. It would be wrong to assert categorically that this constitutes a fault of hierarchical classifications; this is their inherent quality, which renders them higly effective for broad thematic searching when conducted in the conditions of traditional realization in the form of card catalogues. Like other IR languages, hierarchical linear classifications have a restricted range of applications.

Moreover, they are undergoing certain modifications intended to overcome these limitations. Already the UDC offers some facilities for synthesizing notation, which to a certain degree permits recognition of different aspects of a subject classified. Thus, using the relation (colon) sign one can index the multidimensional subject "Bibliography of atomic physics" as 016:539.1 where 016 stands for Special Subject Bibliographies and 539.1 for Atomic Physics. The synthetic notation principle has a very significant role to play in the development of classification.

In this respect, one of the leading figures in classification theory, the British information scientist D.J. Foskett, has this to say: "There is nothing to be gained by discharging hierarchical classification altogether, provided that we recognize that classification, in the modern sense, can and should mean more than this; that it can cope with the most detailed forms of subject analysis.....It is evident that a scheme of terms, whether systematic or alphabetical, however well equipped with cross-references, cannot hope to predict all the contexts in which every term may appear, at some time or another. It can, however, provide a set of roles - operating procedures by means of which such contexts may be freshly created out of the scheme as occasion demands. This means that a classification scheme should no longer set out to provide a 'place' for every document, in the sense that a term, or set of terms, will be found in the actual schedules of the scheme for every subject that may be found in a document. In a modern scheme, the art of the classifier in a library is to construct a symbol that is, in effect, a translation of a subject.(1)

1. Foskett, D.J. Some fundamental aspects of classification as a tool in informatics. In: On Theoretical Problems of Informatics (FID 435), Moscow, 1969, p. 65

Faceted classifications

The first step in this direction was made by the UDC: the introduction of the common subdivisions and special auxiliaries and the connective symbols, in particular the colon, has appreciably enlarged its capacity for building numbers for complex and multi-aspect subjects. But it is to the eminent Indian library scientist S.R. Ranganathan that we owe a fundamental and consistent approach to the solution of this problem. The Colon Classification which he developed in 1933 was a further elaboration of the synthetic principle in classification.

S.R. Ranganathan opposed the practice of building minute 'enumerative' classification tables, in which the compilers sought to provide a separate number for every subject and concept. Instead of a single order to divisions in every main class, he worked out tables each based on one characteristic or aspect, later called 'facets'. The classification number for any document is built from the symbols used in every table, connected by means of the ' colon '. Hence the name - Colon Classification. This classification scheme has not found wide application, but the idea of facet analysis, which forms its most important premise, gave a powerful spur to the development of the library classifications called faceted or analytico-synthetic.

Facet analysis is essentially as follows: first, a field of science or technology is thorougly analysed, a faceted classification scheme is built, and a collection of documents in the relevant field is studied. The analysis yields a list of main facets, while the documents provide the essential terms in the subject field, grouped by the appropriate facets. Each term in a facet is called a focus. A special significance is attached to the order of the foci within a facet, and the facets within a classification scheme. This order is called the facet formula. S.R. Ranganathan suggested a facet formula that includes five categories:

Personality (class, subclass, subject)
Matter (material)
Energy (operation, process, action)
Space (place, territory)
Time.

A number of Indian classification experts and a group of British scientists are presently engaged in the further improvement of this formula, increasing the number of the categories to be included. The indexing procedure in a faceted classification begins with the formulation of the main subject of a document. It is expressed by a chain of foci which are taken from the facets and arranged in a fixed order. Often instead of the foci their numbers will be used. Such a procedure enables classes to be formed for those documents whose subjects are expressed by a combination of characteristics viewed from different aspects. To illustrate this rather complex procedure, we shall present an example from the field of medicine (denoted in Ranganathan's system by the letter L):

Facet O	Facet P	Facet H
Organs of human body	Problems of medicine	Care and Treatment
1 Organism as a whole	1 Preliminaries	1 Nursing
2 Digestive system	2 Morphology	2 Etiology
23 Esophagus	3 Physiology	3 Symptom and diagnosis
24 Stomach	4 Disease	4 Pathology
25 Intestine	42 Infection	
3 Circulatory system	421 Tuberculosis	
4 Respiratory system		
45 Lung		

In classifying documents dealing with the "diagnosis of infectious diseases of the intestine", they will be indexed L 25:42:3, and the"pathology of lung tuberculosis" L 45:421:4.

Compared with hierarchical classifications of the enumerative type, faceted classifications offer a number of attractive features: they greatly facilitate the multi-aspect indexing of documents, bringing together in one place all aspects under which a subject or theme is discussed; they are more hospitable to new terms; they provide greater depth of indexing with shorter notation. However, even the faceted classifications will not ensure searching by any combination of characteristics, as this would necessitate the classified catalogue providing places for all possible groupings and regroupings of the foci taken from the different facets - with card catalogues it is as complicated as in hierarchical classification. Furthermore, the present level of development of faceted schemes permits their effective use only in very specialized document collections.

We have discussed the principal types of conventional IR systems: author, subject and classified. With knowledge of the rules for bibliographical description and the filing of entries in an alphabetical sequence, the techniques of subject cataloguing and classification open the way to mastering IR systems, which have evolved in the course of centuries of development and will for a long time to come serve as large-scale tools of information work. Naturally, the active utilization of these systems requires a deeper understanding of their languages and practical experience in their use, but if you have formed an idea of the merits of these conventional systems and the limitations in their use which are imposed by their structure, it will be easier for you to proceed to the study of the descriptor-type IR systems, which are of fairly recent origin. They will be discussed in the next chapter.

Questions for self-checking

1. What is the purpose of the basic types of conventional IR systems and what are their specific features?
2. In what context do the problems of bibliographical entry occur and how are they resolved?
3. What properties distinguish the IR languages of subject IR systems?
4. What is the structure of a hierarchical library classification?
5. What are the merits and demerits of the UDC?
6. What are the advantages of the faceted schemes over the hierarchical classifications?

Literature

1. Foskett, D.J. Some Fundamental Aspects of Classification as a Tool in Informatics. In: On Theoretical Problems of Informatics (FID 435). Moscow, 1969, p. 64-79.
2. Guide to the Universal Decimal Classification (UDC). B.S. 1000C:1963. (FID No. 345). London, British Standards Institution, 1963, 128 p.
3. Metcalfe, J. Subject Arrangement and Indexing of Information. Notes for Students. Sydney, Bennett, 1966, 181 p.
4. Needham, C.D. Organizing Knowledge in Libraries. An Introduction to Classification and Cataloguing. London, Deutsch, 1964, 259 p.
5. Sayers, W.C.B. A Manual of Classification for Libraries. 4th ed. London, Deutsch, 1967, 404 p.
6. Vickery, B.C. Classification and Indexing in Science. 2nd ed. (enl). London, Butterworths, 1959, XIX, 235 p.
7. Vickery, B.C. Faceted Classification. New Brunswick, Rutgers Univ. Press, 1966.

7. DESCRIPTOR INFORMATION RETRIEVAL SYSTEMS

We have made it clear in the preceding chapter that the conventional information retrieval systems, both subject systems and those based on hierarchic and faceted classifications, along with definite advantages suffer from a certain limitation: they do not provide for document search by any combination of characteristics that has not been established beforehand. The conventional IR systems, particularly alphabetical subject systems and hierarchical classifications, become too cumbersome when confronted with the necessity of providing for multi-aspect search, i.e. searching for documents by multiple characteristics belonging to the different aspects in which a subject or phenomenon is viewed. This limitation becomes even more pronounced when a search for information is carried out in a file of documents that have many such aspects, and the aspects do not fit into a natural hierarchy of generic relations.

An instance of the limitation of conventional IR systems

To remind you of this feature of the conventional IR systems, we shall use as example an ordinary and deliberately simplified case where these systems appear in a schematic form. Let us assume that we conduct an information search in a file of motion pictures according to four aspects containing two characteristics each. Motion pictures are normally divided, according to their content, into documentary and art(non-documentary) films; according to their format, into normal and wide-screen; according to their colour, into black-and-white and colour; according to their length, into short and full-length.

In a *subject system* the name of each of these eight characteristics will form a heading and all of them will be arranged alphabetically:

Art (non-documentary) films	Full-length ....
Black-and-white films	Normal ....
Colour	Short ....
Documentary	Wide-screen ....

The divisions under each of these headings will list the address codes (numbers) or titles of those motion pictures which possess the relevant characteristics. You remember that this system is easy to use, practically does not require any preliminary study by the user of the principles of its build-up, and will quite readily accept new characteristics. For example, if we were to index educational or popular science films, stereoscopic or cinerama films, we would only have to insert into the alphabetic listing the subject headings corresponding to these characteristics. Then in answer to a request for a film possessing any of these characteristics the system would promptly respond with the needed information.

The situation is radically changed, however, when choice of a film by a combination of several characteristics is desired. The system is unable to provide an answer to a multi-aspect query such as this:

"what colour wide-screen art films or films with alternating colour and black-and-white images are available?" To supply an answer, all the relevant subject headings would have to be looked up and the matching film titles identified. The subject system is not designed for this task. We know that a way of overcoming this limitation is to form compound subject headings which combine several characteristics to provide for the would-be requests. However, it is not always possible to foresee all possible requests, and apart-from this, the system then becomes too complex and cumbersome.

Now let us look at a *hierarchic classification*. Taking the first letters of the listed characteristics as the codes of relevant subject headings, and observing the formal rules of construction of hierarchic classification, we shall obtain the classificatory tree presented in Fig.26. This example obviously demonstrates that the hierarchy of generic relations will not at all necessarily be a natural one. Here the choice of the aspects to serve as the characteristic of division at the upper levels of hierarchy depends on the nature of the potential information requests. But they are hard to predict, for they depend both on the purpose of the show, on the type of film-goers, on the time they are ready to spend, and on the projection equipment available.

With the classification chosen in Fig. 26, the system is unable to provide an answer to a multi-aspect request on the availability of short or full-length motion pictures (eight subject headings will have to be scanned) or the black-and-white and colour pictures (four headings will have to be scanned). It is also clear that the necessity of following the formal rules of logic in the construction of a hierarchic classification contributes to the formation of rarely used headings.Thus, according to the above given example, the headings ANBS, ANCS, AWBS and AWCS will not contain very many entries, since short art films are produced infrequently.

To enable the system to give answers to any request using any combination of characteristics, it would be necessary to employ successively each aspect of each characteristics at each level in the hierarchy. This is impracticable because of the very great number of subject headings that would be formed. Another important point to make is that in order to introduce new characteristics, e.g. for the indexing of educational and popular science films or stereoscopic and cinerama films the system would have to be considerably revised to accommodate new classes and subclasses with all the subordinate headings.

Motion pictures where the colour and black-and-white images are alternatingwould be entered under several subclasses at the same time. To enable the system to retrieve only these films, new subclasses uniting both characteristics would have to be formed within it, thereby further increasing its complexity. The system would be made ever more complicated if it were necessary to incorporate into a hierarchical classification of films new principles of division, for instance, according to their country of origin (domestic, foreign), or according to the age group of the film-goers for which they are meant (for children,for adults), and so on. It would result in new levels of hierarchy and many new subject headings.

Faceted classifications are able to eliminate some of these inadequacies. Since in compiling them no attempt is made to list as many combinations of characteristics as possible, but instead a sort of building-block set is offered in the shape of foci (characteristics) grouped in facets (aspects), the indexer can omit practically unessential cha-

racteristics without running the risk of disturbing the hierarchy. He can also form new subject headings of a specific narrow interest. In our example, for indexing short documentary newsreels a generalizing subject heading D:S could be set up, and for art films we could use the subject headings A:N:B; A:N:C; A:W:B; A:W:C. If necessary, to these can be added, leaving the system intact, subject headings for motion pictures with alternating colour and black-and-white images A:N:BC; A:W:BC, as well as subject headings for the new characteristics (foci)- educational films, cinerama films, etc.

It would not be too difficult to insert new aspects of pictures (e.g according to their place of origin, the film-goers age group, etc.)into the faceted classification but it will still not ensure a search for any combination of characteristics, because for this it would have to comprise no fewer subject headings than a hierarchic classification. In our example the system fails to answer a query on all short colour films or black-and-white wide-screen films.

Therefore, the conventional IR systems, which have for centuries been evolving to provide answers to broad thematic and single-aspect requests, have proved to be poorly equipped for specific and multi-aspect searches, and for searches for any combination of characteristics not previously established. The most extensively used conventional IR systems, those based on hierarchical classifications, have great difficulty in catering for the ever increasing number of multidisciplinary problems often without any clear-cut generic relations.

These circumstances led to the emergence, some twenty years ago, of a new method of information retrieval which was to be called *coordinate indexing*. This method forms the basis of the descriptor-type IR systems, discussed in this chapter.

Set-theory terminology

In presenting the essentials of coordinate indexing, it is convenient to use the elementary terms of set theory, some of which it may be useful to introduce here. A *set* is a collection of objects: letters of the alphabet, articles in periodicals, books on shelves, the numbers 1,2,3,4, 5 - each of these is an example of a set, which may contain as little as one element or be empty (contain no elements). Sets may also be infinite, e.g. the infinite number of points on a circle. A set is generally denoted by writing its elements within braces,e.g. $\{1,2,3,4,5\}$. The statement "A is a subset of B" (i.e. every element of A is an element of B) is written in a contracted form as $A \subset B$ or $B \supset A$ (the *inclusion* relation). An empty set is denoted by zero; a set which comprises all the objects in a given field is called universal.

The main operations on sets are the logical sum, logical product, logical difference, and logical complement. The *union* (sum) $A \cup B$ of two sets A and B is a set each element of which is an element of A or of B or of both. The union of the set $\{1,2,3\}$ and the set $\{2,3,4,5\}$ is the set $\{1,2,3,4,5\}$. The *intersection* (product) $A \cap B$ of two sets A and B is a set each element of which is simultaneously an element of A and of B. The intersection of the sets $\{1,2,3\}$ and $\{2,3,4,5\}$ is the set $\{2,3\}$. The *difference*, A - B, between two sets A and B is a set of all elements of A that are not elements of B. The difference of the sets $\{1,2,3\}$ and $\{3,4,5\}$ is the set $\{1,2\}$. The *complement*, A/B, of a set A in a set B is a set of all elements of B that are not elements of A. The complement of the set $\{1,$

2,3} in the set {2,3,4,5} is the set {4,5} . The completement, A or A', of a set A in the universal set is a set of all elements of the universal set that are not elements of A. The graphic representations of the relations between sets are given in Fig. 27.

Coordinate indexing

The coordinate indexing method is based on the assumption that the semantic contents (subject) of a document and information request can with sufficient accuracy and completeness be expressed by an appropriate list of so-called key-words which are explicitly or implicitly contained in the text being indexed. By *keywords* we mean the words which are most essential for expressing the main meaning of a word or phrase, which have a nominative function. Most nouns, adjectives, verbs, adverbs, numerals and pronouns can be used as keywords. Prepositions, conjunctions, connectives, particles and other functional words cannot be keywords. To put it differently, coordinate indexing is a method of expressing the primary subject of a document or information request by a given number of keywords. In pure coordinate indexing, the keywords in the search patterns are not related to each other and function independently. In contrast to search patterns of documents, search request formulations are presented as logical sums, products or complements of the classes designated by the corresponding keywords. In order to locate a document matching an information request it is necessary to do certain logical operations on the classes designated by the keywords in the search patterns of the documents.

In the simplest case, when a search request is formulated as the logical product of a certain set of keywords, the document is held to be a match if its search pattern contains all the keywords of the search request formulation. In a sense, this is equivalent to the operation of logical product of classes designated by those words of the search pattern which coincide with the keywords constituting the search request formulation.

To illustrate this point we shall again take the example given in the beginning of this chapter. Suppose that our collection of films is indexed by keywords denoting the characteristics that have been chosen. To take out of this collection the short films with alternating black-and-white and colour images (a difficult question for conventional IR systems)it will be sufficient to state the request as the logical product of the three keywords describing these characteristics: $S \cap B \cap C$. Another, more abstract example was given in Fig. 16, at the beginning of the fifth chapter, where we conducted a search by means of keywords designated by the letters A to H. We wanted all documents containing characteristics C and D or C and F. Using the terminology that we have introduced we can now say that the request formulation was composed as a logical sum of two products of keywords $(C \cap D) \cup (C \cap F)$.

Externally, the coordinate indexing method appears to resemble an IR language of a subject system. In either case we are dealing with classes designated by appropriate natural-language words and phrases, that is, by keywords and subject headings. Each keyword, e.g. 'Electric', covers a class of subjects and concepts, the designations of which include this word. The same is true of the formulation of subject headings. But the similarity ends there.

In coordinate indexing, in order to locate a document whose search pattern is composed of a given number of keywords, it is necessary to perform the logical operations of sum product or complement on the classes designated by these keywords. This provides an opportunity of conducting a multi-aspect search for any previously unspecified combination of keywords. In the subject systems, on the other hand, as well as in the hierarchical and faceted classifications, each subject heading or classification number appears independently, listing the address codes (numbers) of all relevant subject material. In conventional IR systems simultaneous searching for several subject headings or class numbers is a difficult matter indeed, and sometimes a downright impossibility, as they are not designed for such use. For example, a search for documents using any combination of five mutually exclusive characteristics will require, in a system based on coordinate indexing, a search request formulation as a logical product of five keywords. In any conventional IR system the same task will involve scanning the contents of as many subject headings as there are permutations of five headings of classification numbers, i.e. $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$. This constitutes one of the most important advantages of coordinate indexing over the IR languages of the conventional systems.

Nevertheless the procedure of pure coordinate indexing described above is inadequate since it does not ensure the necessary levels of recall and precision. Some of the main reasons for this are as follows. Natural language words from which keywords for coordinate indexing are chosen have the property of *synonymity:* the same thing or idea may be denoted by different words. Such words which are mutually replaceable in similar contexts without giving a change in meaning are called synonyms. Examples of synonyms are 'test' and 'trial', 'linguistics' and 'science of language', 'common salt' and 'sodium chloride'. It can be easily understood that if the same subject is designated by different words in the search pattern of a document and in a search request formulation, the document, though relevant to the request, will not be retrieved.

Another source of the inadequacy of pure coordinate indexing is the multiple meaning of natural language words. It manifests itself in *polysemy,* which is a transfer of meaning from one subject to another having common qualities or characteristics, and in *homonymity,* which refers to the accidental coincidence of different words that have nothing in common either in respect of origin or meaning. Examples of polysemantic words are 'star' - a celestial body or a geometrical figure, 'bureau' - a kind of table and a collegiate body; examples of homonyms are 'mail' - the system of communication and an armour; 'ray' - a beam of light and a sea-fish. Evidently, the multiple meaning of keywords may bring about the retrieval of non-relevant documents which means lower precision values and increased 'noise'.

One more source of insufficient effectiveness of information retrieval by coordinate indexing may lie in *unspecified relationships* between keywords, which result in the non-retrieval of relevant documents - that is, in poor recall. Let us assume that a search request is for documents on the properties of liquids, and the search request formulation includes the keyword 'liquid': if it is in no way indicated that water is a liquid, the documents dealing with the properties of water will not be retrieved. The most important relations between keywords are the generic relations, but certain significance is attached to other relations,

particularly to the associative relations, i.e. the relations between keywords by association: by contiguity ('chair' and 'table'; 'fall' and 'autumn'), by similarity ('clock', 'scales' and 'thermometer'; 'pond', 'lake' and 'sea'), by contrast ('black' and 'white'; 'light' and 'heavy'), etc. All the conceptual relations between keywords, based on the relations between the ideas, things or phenomena they designate, are called paradigmatic relations.

Descriptor information retrieval language

It follows from the examples discussed above that, in order to effect a considerable increase in the precision and recall of information retrieval based on coordinate indexing, it is necessary at least to establish control over the vocabulary of the IR language used. Such a control should consist in the elimination of synonymity and in compiling special dictionaries, tables or charts graphically demonstrating the most essential paradigmatic relations between the keywords. With full vocabulary control, the coordinate indexing of search requests and documents will make use of only those words contained in a given standard list, in which their synonymity and polysemy has been eliminated and the paradigmatic relations are shown. Such keywords have been termed 'descriptors'.

Descriptors are standardized keywords, designed for coordinate indexing of documents and information requests, selected according to certain rules from the main vocabulary of a particular natural language and artificially (by cross-references and notes) freed from synonymity, polysemy and homonymity. A *descriptor language* is a specialized IR language, whose vocabulary is composed of descriptors and whose grammar consists, in the simplest case, of a method of building search patterns and search request formulations by correlating relevant descriptors.

It is impossible to establish now who was the originator of the idea of coordinate indexing. The British information scientist B. Vickery has noted that coordinate indexing with inverted file structure was apparently known in Sumer three thousand years ago. Clay tablets have been found, each assigned to a disease, with the names of all diseases characterized by a particular symptom also listed on the relevant tablet. Obviously, if according to the symptoms observed in a patient the corresponding tablets are selected and the name of any disease repeated on each of them is noted, it will in all probability be the disease from which the patient in question is suffering. These Sumerian tablets are the prototype of modern diagnostic machines. (1)

The principle of coordinate indexing is the basis of the so-called superimposable cards, which will be considered in the next chapter. In 1915, they were used by the ornithologist H. Taylor, in the 1920's they found applications in mineral identification and personnel selection, and in 1939 W. Batten (of Imperial Chemical Industries, Great Britain) used them for patent searching. The credit for developing the information retrieval systems which formed the basis for the development of modern descriptor languages, however, must go to the American scientists C.N. Mooers and M. Taube.

1. International Study Conference on Classification for Information Retrieval. Dorking, 1957. Proceedings. London, Aslib, 1957, p. 106.

C.N. Mooers in 1947 designed a mechanized document searching device which he named "Zatocoding system", and it was he who coined many of the terms which are now in wide use among information scientists, including 'information retrieval', 'information retrieval system', 'retrieval language ', 'descriptor' and 'descriptor dictionary'. In one of his earlier works, he has described the indexing method in his system thus: "Each document or unit of information is characterized by a set of descriptors taken from the vocabulary of descriptors. Each descriptor of the set applies to, or is true in some way, of the information content of the unit of information. The descriptors operate independently in this type of characterization. The fact that there are several descriptors in the set may mean that they formed some interacting combination in the original document, or it could just as well mean that they relate to independent ideas scattered through the document. Using descriptors in this fashion drops almost all relationship between the ideas represented by the descriptors.(1)

Mr. Taube made a great contribution to the theoretical substantiation, development and popularization of the ideas of coordinate indexing. He defined coordinate indexing as a 'method of analysing items of information so that retrieval is performed by the logical operations of the product, sum and complement on the codes in the store'(2). In 1951 he developed the socalled Uniterm system which has found extensive application since then. The main difference between the Uniterm and Zatocoding systems is that the former deals with words expressing concepts while the latter deals with concepts expressed by words.

Uniterm system

Taube's Uniterm is a keyword (usually a simple one) which may have an appropriate cross-reference or scope-note helping to eliminate its synonymity or multiple meaning. In contrast to the descriptors of C. Mooers, the Uniterms have no references specifying the paradigmatic relations between them. Uniterms may be keywords expressing single ideas, as well as proper, geographical and trade names. All Uniterms have an equal hierarchical rank: none of them occupies a conceptually superior position with respect to any other Uniterm (as in hierarchical classifications) and none is used in a pre-established combination with any other Uniterm (as in subject headings).

This method of building an IR language vocabulary contributes to a great reduction in its size. The list of subject headings in the Library of Congress subject catalogue contained some 50,000 entries; conversion to Uniterms resulted in a list containing only 3,000 words. The important feature of the Uniterm system is that the vocabulary is built in the process of using it and not developed beforehand, as in the Zatocoding system. During the initial period of operating the system, the number of Uniterms will grow steadily. Gradually, synonyms and polysemantic words will be detected and marked on Uniterm-cards. With the growth of the document collection, the rate of growth of the vocabulary will tend to slow down and eventually come almost to a standstill.

---

1. Mooers, C.N. Zatocoding Applied to Mechanical Organization of Knowledge. "American Documentation", 1951, V. 2, No. 1, p.26.

2. Taube, M., Wooster, H. (eds.). Information Storage and Retrieval. Theory, Systems and Devices. New York, Columbia Univ. Press, 1958, p.8.

Using as an example this system, one of the first and simplest systems based on the coordinate indexing method, it will be easy to become familiar with the practical techniques of input and searching in a descriptor-type IR system. For the physical implementation of his system, Taube developed a special Uniterm card having either the usual catalogue card size (75mm x 125mm) or larger in size (203mm x 125mm), on which a grid formed by one horizontal and ten vertical lines is printed. The horizontal line is used for writing a Uniterm. Uniterm cards are envisaged for all Uniterms in the system and are filed alphabetically, with due account taken of the inversion of those Uniterms consisting of more than one word.

The vertical columns are for the codes (numbers) of documents whose search patterns include the Uniterm indicated in the horizontal line of the card. The peculiarity of the system is the "terminal digit" order of writing these numbers on the card. This means, for example, that number 127 is written in the 7 column, number 239 in the 9 column, and number 270 in the 0 column. Since the documents are numbered as they are entered into the system, the numbers will be posted in the columns in ascending order. Such a system of recording considerably facilitates the visual scanning and locating of document numbers which are common to all the cards being matched. The procedure in this case amounts to the operation of logical multiplication of the Uniterms contained in a search request formulation. The method of work with Uniterm cards is demonstrated in Fig. 28.

Let us assume that we are to enter into the system a paper dealing with methods of 'corrosion protection of gas turbine blades'. For the coordinate indexing of this paper we need the Uniterms TURBINES, GAS, BLADES, CORROSION and PROTECTION, which jointly form its search pattern. The paper is assigned its serial address code 2005. Then the cards assigned to these Uniterms are removed from the file. If in doing this, the Uniterm PROTECTION or its synonyms were found not to be in the system, a card would be made for this particular Uniterm. The number 2005 is put in the 5 column of each of the relevant Uniterm cards, which are then replaced in the file. Similar procedures are performed for each incoming document.

An information search in a file of Uniterms is conducted in the following way. Suppose that a request is put to the system for papers dealing with 'gas turbine blades'. The search request formulation is made to contain the Uniterms GAS, TURBINES, BLADES. The cards for these features are extracted from the Uniterm file and the numbers common to all these cards are located. It is advisable to have as the basis for comparison the card which contains the least number of postings (in this case, the Uniterm card for BLADES). Matching should likewise begin with the columns with the least number of postings: columns that are empty in any of the cards need not be matched at all; where the base card has columns with one posting it is not necessary to match the postings which have been made before this single posting (i.e. smaller numbers). The time it takes to locate all common postings is usually negligible. Thus, in our example, it can be seen at once that the three upper Uniterm cards have only three matching numbers: 526, 1027 and 2005, which are the address codes of the documents sought. If the request is broadened to cover documents on gas turbines (two cards at the top), the output will increase to include five documents: 195, 294, 526, 1027, 2005. If, on the other hand, the request is made more specific to read "corrosion/of gas turbine

blades" (four upper Uniterm cards) only two relevant documents will be retrieved, namely 1027 and 2005.

Thesaurus and its construction

In the Uniterm system, the control over the retrieval language vocabulary, as we have seen, is limited to the elimination of synonymity and polysemy. Full lexical control, however, which alone can ensure maximum precision and recall, especially with thematic searches and searches for documents only partially matching the request, necessitates the recognition of paradigmatic relations between the indexing terms. For this purpose special normative reference dictionaries called thesauri are compiled.

An information retrieval *thesaurus* is a reference dictionary designed to help the information user to state his information needs in terms of the descriptor language and to provide for finely detailed indexing of documents and information requests by these terms. It must contain all the descriptors used by the language of a given system, clearly displaying their conceptual relationships, and also the key-words within the system which are considered to be synonymous with these descriptors. Thus, the thesaurus will help to eliminate synonymity and polysemy of keywords, and the lack of explicit relations between them, which give rise to the defects of 'pure' coordinate indexing already discussed.

Synonymity of keywords is eliminated in the following way. In constructing a thesaurus, a list of keywords used for the coordinate indexing of documents is first compiled. Then groups of words which can be considered synonyms are brought together. From each of these groups a word or phrase is taken to represent the whole group, which becomes a descriptor. The rest of the words in the group are considered synonyms of the descriptor and linked to it by 'see' or 'use' references. Each descriptor is linked with all its synonyms by reversed 'includes' or 'used for' references, e.g.

ABSTRACT includes	speculative theoretical
speculative see	ABSTRACT
theoretical see	ABSTRACT.

Polysemy of keywords is also provided for, in building up a dictionary of descriptors, by affixing, to each multiple-meaning keyword an alphabetic or numerical symbol and a word qualifying its meaning. Another method is to replace polysemantic keywords by descriptors composed of one-value phrases, e.g.:

atlas(geographical) see geographical atlas
atlas(vertebra) see cervical atlas
filter - 1 (chemical)
filter - 2 (electrical)
filter - 3 (gas)
filter - 4 (optical)

The conceptual (paradigmatic) relations between discriptors are also displayed in the thesaurus: according to our definition , these are

relations based on the existence of objective connections between the ideas, objects or phenomena denoted by the descriptors. In this case synonimity is not taken into account, because in a descriptor language it is eliminated and does not exist within one system. The following paradigmatic relations can be cited as the most important ones:

species - genus (genus - species);
collateral subordination;
similarity (functional);
cause - effect (effect - cause);
part - whole (whole - part).

Of particular importance are the generic relations and the collateral subordination relations realizable through them. It is these relation types that form the groundwork of hierarchical classifications. Almost every concept can be both generic (i.e. it can reflect major features of a class of objects that includes other classes of objects which are species of this genus) and at the same time specific. The notion of 'rectangular', for example, is generic to the notion 'square' and specific to the notion 'parallelogram'. Exceptions are constituted only by the broadest concepts or categories (e.g. 'matter', 'space', 'time'), which have no generic concepts, and by the most narrow, unique notions (e.g. 'UDC') which have no specific concepts. Concepts which are equally subordinated to one generic concept are said to be collaterally subordinated, e.g. the concepts 'phonetics','lexicology' and 'grammar' are collaterally subordinate to the generic concept 'linguistics'. It should be noted that, in contrast to hierarchical classifications, descriptor languages take into consideration generic relations irrespective of the hierarchic level to which they belong.

A thesaurus indicates generic relations independently of other paradigmatic relations which may sometimes outwardly resemble them. This applies to the functional similarity relations (clock - scales - thermometer), causal relations (tiredness - sleep), and particularly to the relations of the 'part - whole' kind: of all these, it can be said that they differ from generic relations in that they represent relations between objects and not between concepts; 'genus' and 'species' are abstractions, while 'whole' and 'part' are concrete things.

Thus, an information retrieval thesaurus generally consists of three parts:

1. The vocabulary part is a normal alphabetic list of descriptors, together with keywords regarded in this system as synonyms of these descriptors. Descriptors are commonly made more prominent in the listing (e.g. by printing them in capitals) and are linked by cross-references to and from all their synonyms. Polysemy and synonymity of keywords are eliminated using the method described above.

2. The 'semantic map' of the retrieval language vocabulary is a network of conceptual classes in which all the descriptors of a given language are grouped. This part of the thesaurus provides a graphic demonstration of the essential paradigmatic relations between descriptors, at least of their generic relationships. These relations are expressed in one of two ways: either by combinations of the alphabetically listed thematic classes (fields) containing a multiplicity of thematic groups of descriptors also listed alphabetically, or by charts in which the basic paradigmatic relations are indicated by arrows.

3. The rules of conversion of keywords and key phrases of the natural language into a descriptor-type IR language determine the procedure of substitution of descriptors for these keywords and phrases. The rules define the conversion of the names of institutions, chemical compounds, biological species, and other similar categories of terms, and also include the rules for lexicographical editing of search patterns and search request formulations, e.g. rules for complementing them with descriptors connected with the main descriptors by generic and other paradigmatic relations.

To illustrate the thesaurus structure two examples will be given. The first published thesaurus to use the graphic method of displaying paradigmatic relations between descriptors was the "Euratom-Thesaurus" (first edition) for nuclear physics and engineering.(1) It contains 4,470 descriptors, including 1,836 names of inorganic chemical compounds and 1,404 names of isotopes. The 42 subject classes, numbered 00 to 94 (with lacunae) are in the form of charts in which the arrows are directed from the generic to specific descriptors. The collaterally subordinate descriptors are marked by two-way arrows. These classes cannot be considered as classifications; they broadly correspond to those subject fields that are of interest to Euratom. The vocabulary part provides a common alphabetical list of descriptors with references to the relevant subject classes, and a common alphabetical list of their synonyms which are associated with the corresponding descriptors and their subject classes. The generic relations are thus completely ignored in the vocabulary part. Fig. 29 shows the subject class 71 'Mathematics' from the "Euratom-Thesaurus" (first edition). The arrows which go beyond the limits of a subject class join its descriptors with the descriptors of the other subject classes and their numbers. To the right, there is an alphabetical listing of all descriptors which belong to this particular class, and at the bottom there are numbers of all subject classes used in the thesaurus.

Another example is the "Thesaurus of Engineering Terms" which was published in 1964 by the Engineers Joint Council (USA) (2). In contrast to the Euratom thesaurus, it indicates generic and other paradigmatic relations by means of a system of references and scope notes. The thesaurus contains 10,515 words of which 7,750 are descriptors. The listing of the descriptors and their synonyms in the vocabulary part is alphabetical. The synonyms are linked with the descriptors meant to replace them by 'Use' references. In the 'semantic map', which has the form of an alphabetical listing of the basic (heading) descriptors, each entry includes the heading descriptor, its synonyms, the specific descriptors, the generic descriptors and the descriptors connected with it by other paradigmatic relations. To designate them the following notes are employed:

UF (used for) - for synonyms;
BT (broader term) - for generic descriptors;
NT (narrower term)- for specific descriptors;
RT (related term) - for other descriptors.

1. Euratom-Thesaurus, Keywords used within Euratom's Nuclear Energy Documentation Project. EUR 500.e (1st ed.) Brussels, 1964, 80 p. (European Atomic Energy Community).

2. Thesaurus of Engineering Terms: a list of engineering terms and their relationship for use in vocabulary control, in indexing and retrieving engineering information. 1st ed. New York, Engineers Joint Council, 1964.

The ampersand (&) is used to mark those descriptors which can be used in place of other descriptors expressing more narrow concepts and marked by the symbol # .

The second major thesaurus published by the Engineers Joint Council, "Thesaurus of Engineering and Scientific Terms", 1st ed. (New York, 1967), uses some other symbols. The dagger (†) in front of a term signifies that two or more descriptors are to be used in coordination for that term. The dash (-) symbol in front of a descriptor indicates that the descriptor has narrower terms (not shown) and that the main entry should be consulted to determine these. Below is a simple dictionary entry from the "Thesaurus of Engineering and Scientific Terms":

Scientists 0509

UF Scientific personnel
BT Personnel
Professional personnel
NT Chemists
Physicists
RT Engineers

A few words about the term 'thesaurus' (from Greek 'thesauros', literally meaning a treasure, treasury or storehouse) might be appropriate The word seems to have been used first in its present meaning by Florentine Brunetto Latini (1220-1294) in his encyclopaedia "Li Livres dou Trésor". In the 16th century, the word appeared in the names of the Latin and Greek lexicons "Dictionarium, seu Lingue Latinae Thesaurus" (1532) and "Thesaurus Linguae Graecae" (1572), which were published by the Estiennes, well-known French philologists and publishers. In contemporary usage the term refers to the dictionaries of concepts, or ideological dictionaries, which are in effect inverted explanatory dictionaries. Of these, the greatest popularity is enjoyed by the "Thesaurus of English Words and Phrases", compiled by P.M. Roget in 1852 and since reprinted at least 90 times. Similar dictionaries, linking all words in thematic groups or subject classes, are to be found in French, German, Spanish and other languages.

One of the first to recognize the need for this type of dictionary in information retrieval work was the American bibliographer C.L. Bernier who wrote in 1957: "A limited thesaurus would seem to be another effective way of bringing the relevant terms to the attention of the searcher if the vocabulary proves too large to be read completely each time for selection". (1)

Grammatical resources of descriptor IR languages

In using descriptor languages, the search pattern of every document and search request formulation is stated in the form of an unordered set of descriptors. In addition to their paradigmatic relations, however, which in some way or other are recognized in the indexing and in searching assisted by a thesaurus, there exist different relationships between descriptors that derive from the document contexts. Such relations between the descriptors are said to be *syntagmatic*. If these re-

1. Bernier, C.L. Correlative Indexes. II: Correlative Trope Indexes. "American Documentation", 1957, V. 8, No. 1, p. 48.

lations are ignored, the descriptors belonging to a document search pattern may form false combinations or 'false drops' which result in the retrieval of non-relevant documents, increased 'noise' and hence lower precision.

The following examples will illustrate this point. Suppose that our system includes a document on the 'production of sulfuric acid and catalyst purification'; its search pattern will contain the following descriptors - PRODUCTION, SULFURIC ACID, CATALYST and PURIFICATION. These descriptors may, during a search, form false combinations which will result in the retrieval of this document in answer to a request on the 'purification of sulfuric acid' and on the 'production of catalyst', although the document does not deal with either of these. (1) Another example: a document dealing with the 'coating of copper tubes with lead' is indexed by the following descriptors - LEAD, COATING, COPPER, TUBES. If a request for documents on lead tubes is entered into the system, the request formulation will be indexed by the descriptors LEAD and TUBES, and the document will be retrieved, although it is not relevant to the query. (2)

Roles (role indicators) and links are the principal grammatical resources used in descriptor languages to reduce noise. *Roles* are special symbols which are attached to a descriptor and reduce the scope of the concept it stands for. This is achieved by indicating the logical role which a given descriptor plays in a particular context. In the first of the examples given, in order to prevent false combinations it will be enough to affix to the descriptor SULFURIC ACID in the search pattern the role indicator "B" showing that it is a product of chemical or industrial reaction, and to the descriptor CATALYST the link "C" indicating an undesirable component (waste, impurity, admixture, spoilage). The document search pattern will then be indexed - PRODUCTION, SULFURIC ACID-B, CATALYST-C, PURIFICATION, whilst the request formulation will read - CATALYST-B, PRODUCTION, SULFURIC ACID-C, PURIFICATION. The document will then not be retrieved in response to the queries asking for documents on catalyst production and sulfuric acid purification.

Links are also special symbols attached to the descriptors in the search patterns of documents (or their address codes) and designed for the conceptual grouping of these descriptors. In the second of our examples , the search pattern of the document on the 'coating of copper tubes with lead' would assume this form by the addition of links - LEAD R_1, COATING R_1, COPPER R_2, TUBES R_2. The link R_1 would join the descriptors LEAD and COATING, and the link R_2 the descriptors COPPER and TUBES. If the condition is stipulated, that a document is to be produced on request only when the matching descriptors in its search pattern and in the search request formulation have the same links ascribed to them, this document will not be retrieved in answer to a query on 'lead tubes'. But then it would also not be retrieved in response to a request for documents dealing with 'tube coatings' generally. This would mean an information loss. The search pattern can be made more complex by the addition of several links: LEAD R_1, COATING R_1R_3, COPPER R_2R_3, TUBES R_2R_3.

1. Holm, B.E. Information Retrieval - a Solution. "Chemical Engineering Progress", 1961, V. 57, No. 8, p. 74-75.
2. Taube, M. Notes on the Use of Roles and Links in Coordinate Indexing. "American Documentation", 1961, V. 12, No. 2, p. 98-100.

If now the condition stipulates at least one common link in the matching descriptors, the document will be retrieved in response to questions on 'tube coating', 'lead coating' and 'coating of copper', but will not be retrieved by a false correlation in response to a request for 'lead tubes'

Experience indicates that the best results are produced by the joint utilization of links and roles. They reduce noise by 10-15%, but considerably increase the cost and time of indexing. In relatively small document collections (up to 30,000 items) the increased precision of retrieval achieved by such grammatical means will not balance the increased time and cost of indexing and searching, which means that under certain circumstances the use of descriptor languages without grammar is more advantageous.

Further development of grammatical means of the descriptor languages has resulted in specialized IR languages in which the semantic relations between descriptors are developed to a greater extent. These relations and the descriptors themselves are designated by complex codes and the search pattern of a document has the form of an encoded 'telegraphic abstract'. In one such language the notion 'thermometer' is treated as a machine or device(MACH.) which effects (U) measurement (MUSR) and is affected (W) by heat (RWHT.) and is designated by the descriptor MACH. MUSR.RWHT. Among the best-known of such languages are the WRU Semantic Code, SYNTOL, the language of RX-codes and a few others. They are still in the experimental stage and designed primarily for automating information retrieval with the aid of computers. Familiarization with these languages is beyond the scope of our introductory course.

We have now considered some of the fundamental problems relating to descriptor-type IR systems; the reasons for their appearance and their advantages over conventional systems; the principles of coordinate indexing and its inherent weaknesses; the basic concepts of descriptor languages; the operation of some of the simpler systems based on these languages; the construction of the thesaurus and its range of applications and, finally, the grammatical resources of descriptor IR languages. To conclude the discussion of information retrieval problems we have only to consider the technical means of implementing IR systems, which will be the subject of the next chapter.

Questions for self-checking

1. How do the inherent limitations of conventional information retrieval systems manifest themselves?
2. What is coordinate indexing and what are its advantages over the IR languages of conventional systems?
3. What are the definitions of a descriptor language and of its principal concepts?
4. How is a Uniterm system designed and how does it work?
5. What is a thesaurus, its structure and purpose?
6. What is the function of the grammatical resources of a descriptor language?

Literature

1. Costello, J.C. Training Manual and Workbook for use in Abstracting and Coordinate Indexing. Training course. Columbus, Battelle Memorial Institute, 1964, IX. 117, 11 p.
2. Howerton, P.W. (ed.). Information Handling: First principles. Washington, Spartan Books, 1963.
3. Lancaster, F.W. Information Retrieval Systems, Characteristics, Testing and Evaluation. New York (a.o.), Wiley, 1968, XIV, 222 p.
4. Taube, M. and Wooster, H. (eds.). Information Storage and Retrieval. Theory, Systems and Devices. New York, Columbia Univ. Press, 1958.
5. Vickery, B.C. On Retrieval System Theory. 2nd ed. London, Butterworths, 1965, XII , 191 p.

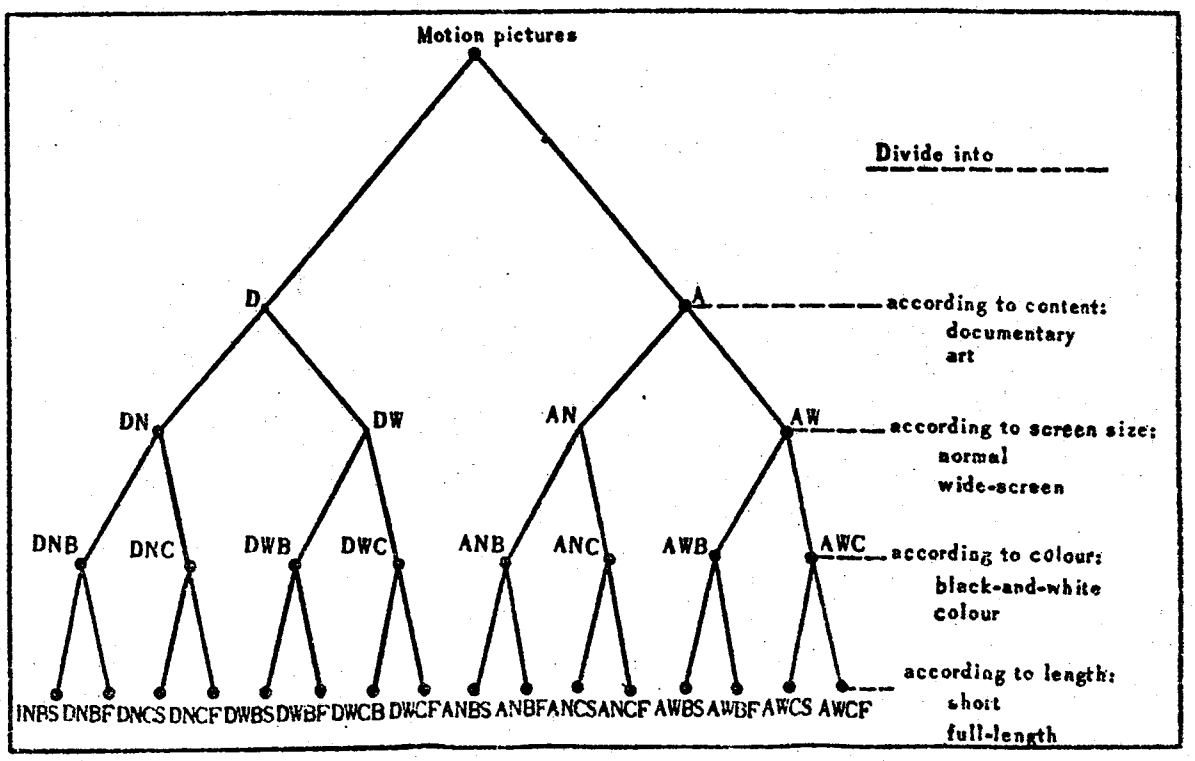


Fig. 26. A hypothetical classification of objects according to attributes which do not form a natural hierarchy.

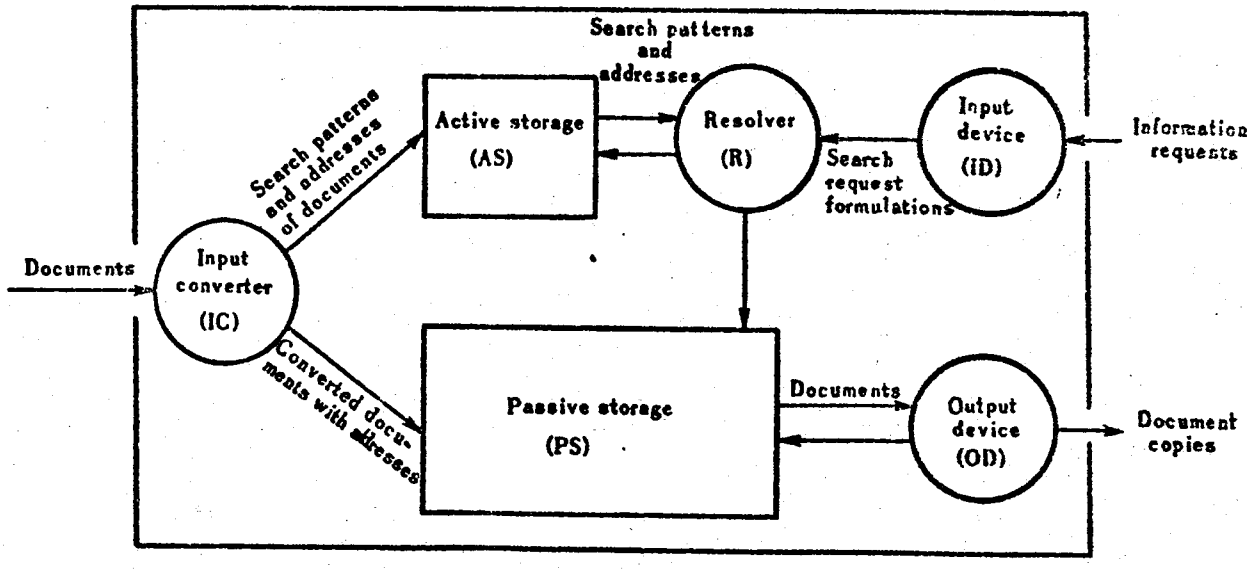


Fig. 18. General block-diagram of an information retrieval system.

Based on the paper, by G.E. Vleduts, "O nekotorykh storonakh issledovaniy po sozdaniyu informatsionno-poiskovykh sistem" (Some aspects of research in the development of information retrieval systems) in "Nauchno-Tekhnicheskaya Informatsiya", 1961, No. 1, p. 32.

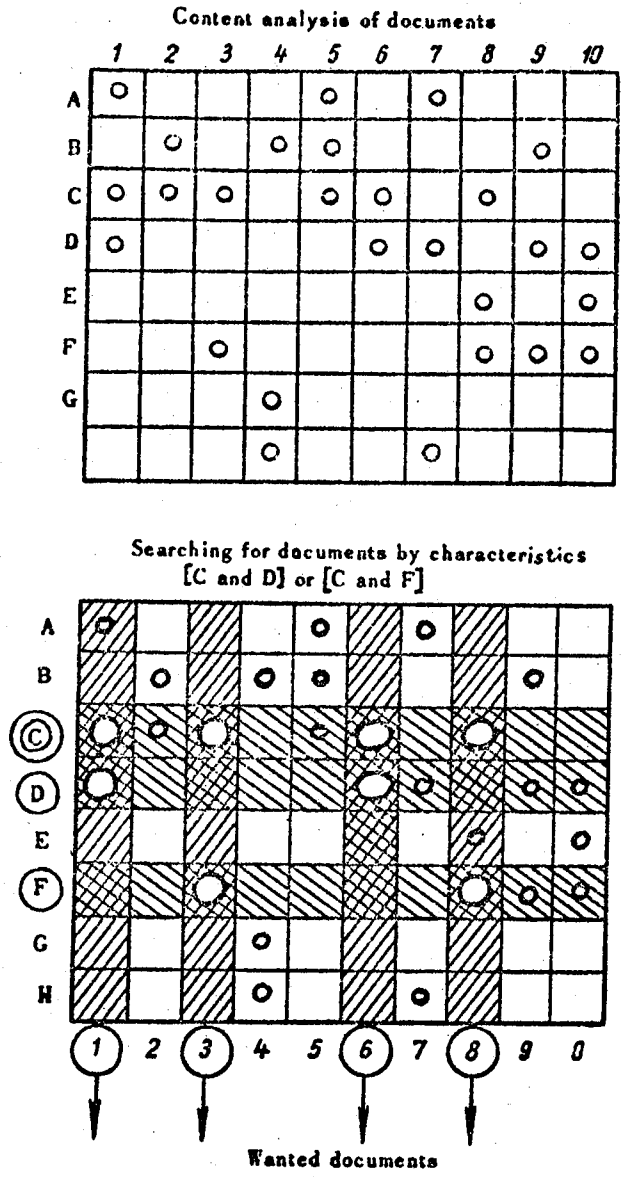


Fig. 16. Diagram of information retrieval.

Used on the paper by B.-A. Lipetz "Information storage and retrieval" in "Scientific American", 1966, vol. 215, No. 3, p. 226.

Documents	Relevant	Non-relevant	
Retrieved	a	b	$a+b$
Non-retrieved	c	d	$c+d$
	$a+c$	$b+d$	

Recall $\frac{a}{a+c} \cdot 100\%$

Precision $\frac{a}{a+b} \cdot 100\%$

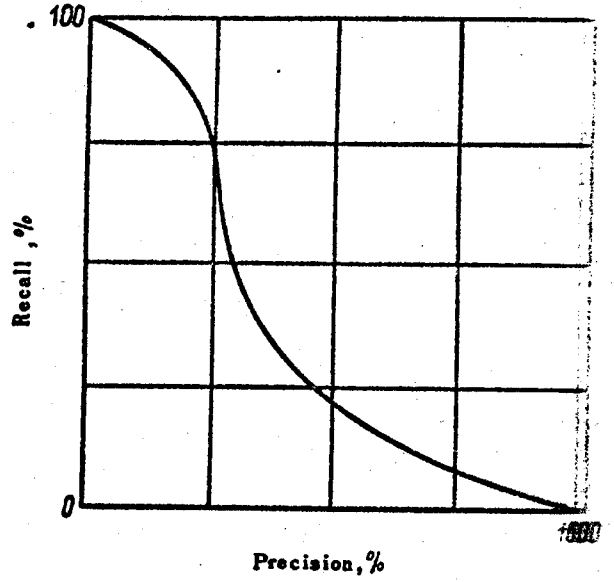


Fig. 17. Dependence of recall on precision (after C. Cleverdon).

C.W. Cleverdon and J. Mills. The testing of index language devices. - "Aslib Proceedings", 1963, v. 15, No. 4, p. 106-130.